BIENNIAL REPORT

2006 – 2007

Laboratory of Computer and Information Science

Adaptive Informatics Research Centre

Helsinki University of Technology

P.O. Box 5400

FI-02015 HUT, Finland

K. Raivio, R. Vigário, and L. Koivisto, editors

Otaniemi, April 2008

ISSN 1795-5092 (Print)
ISSN 1796-4121 (Online)

Yliopistopaino
Helsinki 2008

Contents

Preface

The Laboratory of Computer and Information Science (CIS, informaatiotekniikan laboratorio) was one of the research and teaching units of the Department of Computer Science and Engineering at Helsinki University of Technology until the end of year 2007, after which the organization of The University was radically changed. The CIS laboratory had its roots in the Electronics Laboratory, established in 1965 by Professor Teuvo Kohonen. For more than 30 years, the research in the laboratory was concentrated on neurocomputing, especially associative memories, self-organization, and adaptive signal and image processing, as well as on their applications on pattern recognition. The laboratory grew from its roots of one professor and a handful of students and researchers into a relatively large and well established unit of 5 professors and about 80 staff altogether.

For 12 years, from 1994 to the end of year 2005, the majority of research in the laboratory was carried out within the **Neural Networks Research Centre (NNRC, neuroverkkojen tutkimusyksikkö)**, established by Professor Kohonen as a separate research unit with its own funding and own administrative position. It was selected as one of the first Finnish national Centers of Excellence in Research in 1995. The Academy of Finland extended its Center of Excellence status for the years 2000 to 2005.

From January 2006, the activities of the NNRC were inherited by a new and larger research unit, whose name **Adaptive Informatics Research Centre (AIRC, adaptiivisen informatiikan tutkimusyksikkö)** reflects the changing emphasis of the research conducted by us. This unit, too, was selected by the Academy of Finland as a national Center of Excellence for the years 2006 - 2011 after a very tough competition. This status has also implied financial resources from the Academy, Tekes, HUT, and Nokia Co., which are gratefully acknowledged.

During 2006 - 2007, the AIRC was operating within the Laboratory of Computer and Information Science, coordinating the major part of its research activities. It is not possible to separate the personnels of these two units, as the teaching staff of the CIS also participated in research projects of the AIRC. Professor Erkki Oja was both the director of AIRC and the head of the CIS laboratory. Professor Samuel Kaski was the vice-chair of AIRC, with Professors Olli Simula and Juha Karhunen participating in its research projects. In addition, 18 post-doctoral researchers, 33 graduate students, and a number of undergraduate students were working in the AIRC projects in 2007.

Professor Heikki Mannila of the CIS Laboratory was partner and vice-director of yet another research centre, the **From Data to Knowledge research unit (FDK, datasta tietoon - tutkimusyksikkö)**, a joint effort between Helsinki University of Technology and the University of Helsinki. Also this research group was a national Center of Excellence during 2002 - 2007. Although the Adaptive Informatics Research Centre and the From Data to Knowledge research unit were financially separate and stemmed from different research traditions, there has been an overlap in the research directions and projects between these two Centers of Excellence. This overlap has produced fruitful joint research which is expected to increase in the future.

The present report covers the activities during the years 2006 and 2007. Basically, the report is divided in two parts. In the first part, the research of the AIRC is reviewed. In the second part, those projects of the FDK research unit are reviewed, that pertain to the research activities in the CIS laboratory. The main reason for this separation is that the present booklet also serves as the official report of the AIRC to its sponsors, and it is important to clearly distinguish exactly what work has been done under those finances.

The achievements and developments of AIRC's predecessor Neural Network Research Centre have been thoroughly explained in the triennial reports 1994 - 1996 and 1997 - 1999 for the first six years of CoE status, as well as the biennial reports 2000 - 2001, 2002 - 2003, and 2004 - 2005 for the second six year period. The web pages of the laboratory, `http://www.cis.hut.fi/` also contain up-to-date texts.

To briefly list the main numerical achievements of the period 2006 - 2007, the laboratory produced 12 D.Sc. (Eng.) degrees, 1 Lic.Tech. degree, and 37 M.Sc. (Eng.) degrees. The number of scientific publications appearing during the period was 234, of which 46 were journal papers. It can be also seen that the impact of our research is clearly increasing, measured by the citation numbers to our previously published papers and books, as well as the number of users of our public domain software packages.

A large number of talks, some of them plenary and invited, were given by our staff in the major conferences in our research field. We had several foreign visitors participating in our research, and our own researchers made visits to universities and research institutes abroad. The research staff were active in international organizations, editorial boards of journals, and conference committees. Also, some prices and honours, both national and international, were granted to members of our staff.

The first meeting of the Scientific Advisory Board of AIRC was held on May 10 - 11, 2007. The evaluation report written by the members of the Board, Professors Risto Miikkulainen and José C. Príncipe, was quite positive. Another evaluation for the whole CIS laboratory was carried out in the context of the Academy of Finland evaluation of Computer Science research in Finland 2000 - 2006. In that evaluation, the foreign experts state e.g. that "The Laboratory has generated outstanding research in its areas and is considered among the world's top ... it has a consistent outstanding publication record at top conferences and journals with spectacular citation impact."

Erkki Oja	*Samuel Kaski*	*Heikki Mannila*
Professor	Professor	Academy Professor
Director,	Vice-Director,	Vice Director,
Adaptive Informatics	Adaptive Informatics	From Data to Knowledge
Research Centre	Research Centre	Research Unit

Personnel

Employees during 2006 – 2007

Professors

Erkki Oja, D.Sc. (Tech.). Director, Adaptive Informatics Research Centre. Director, Laboratory of Computer and Information Science
Olli Simula, D.Sc. (Tech.). Head of Department of Computer Science and Engineering
Heikki Mannila, D.Phil. Vice Director, From Data to Knowledge research unit
Samuel Kaski, D.Sc. (Tech.). Vice-Director, Adaptive Informatics Research Centre
Juha Karhunen, D.Sc. (Tech.), part-time
Teuvo Kohonen, D.Sc. (Tech.), Emeritus Professor, Academician

Post-doc researchers

Francesco Corona, PhD, from Jan. 2007
Gemma Garriga, PhD, from Oct. 2006
Robert Gwadera, PhD, until Jul. 2007
Jaakko Hollmén, D.Sc. (Tech.), Chief Research Scientist
Antti Honkela, D.Sc. (Tech.)
Timo Honkela, D.Phil., Chief Research Scientist
Alexander Ilin, D.Sc. (Tech.)
Mika Inki, D.Sc. (Tech.), until Jun. 2006
Markus Koskela, D.Sc. (Tech.), visiting abroad from Nov. 2005 to Jan. 2007
Mikko Kurimo, D.Sc. (Tech.)
Ville Könönen, D.Sc. (Tech.), until Mar. 2007
Jorma Laaksonen, D.Sc. (Tech.)
Krista Lagus, D.Sc. (Tech.)
Amaury Lendasse, PhD
Janne Nikkilä, D.Sc. (Tech.)
Petteri Pajunen, D.Sc. (Tech.), until Jul. 2007
Jussi Pakkanen, until Aug. 2007
Kalle Palomäki, D.Sc. (Tech.)
Jaakko Peltonen, D.Sc. (Tech.)
Kai Puolamäki, D.Phil.
Tapani Raiko, D.Sc. (Tech.)
Kimmo Raivio, D.Sc. (Tech.)
Jouni Seppänen, until Oct. 2007
Miki Sirola, D.Sc. (Tech.), laboratory engineer
Jaakko Särelä, D.Sc. (Tech.), until end of 2006
Ricardo Vigário, D.Sc. (Tech.)

Post-graduate researchers

Antti Ajanki
Matti Aksela, until May 2007
Jose Caldas
Mathias Creutz, on leave from May 2007
Nicolau Gonçalves
Ramunas Girdziusas
Sami Hanhijärvi
Markus Harva
Hannes Heikinheimo
Heli Hiisilä, until Aug. 2007
Teemu Hirsimäki
Ilkka Huopaniemi
Yongnan Ji, until Oct. 2006
Arto Klami
Mikaela Klami
Oskar Kohonen
Mikko Korpela
Leo Lahti
Pasi Lehtimäki
Elia Liitiäinen
Tiina Lindh-Knuutila
Yoan Miché
Mikko Multanen, until Oct. 2007
Hannes Muurinen, until May 2007
Merja Oja
Jukka Parviainen
Mari-Sanna Paukkeri
Janne Pylkkönen
Matti Pöllä
Juha Raitio
Ulpu Remes
Nima Reyhani
Salla Ruosaari
Jarkko Salojärvi
Eerika Savia
Timo Similä, until Aug. 2007
Vesa Siivola, until May 2007
Mats Sjöberg
Antti Sorjamaa
Mika Sulkava
Nikolai Tatti
Jarkko Tikka
Janne Toivola
Matti Tornio, until Jun. 2007
Ville Turunen
Antti Ukkonen
Matti Varjokallio
Jarkko Venna, until Jan. 2007

Ville Viitaniemi
Sami Virpioja
Niko Vuokko
Jaakko Väyrynen
Zhirong Yang
Jarkko Ylipaavalniemi
Zhijian Yuan, until Sep. 2007

Under-graduate researchers (full-time or part-time)

Tuomas Alhonnoro
Otto Eirola
Andrey Ermolov
Leo Gillberg
Kevin Hynnä, until Mar. 2006
Heikki Kallasjoki
Jakke Kulovesi
Tuomas Kärnä
Golan Lampi
Lasse Lindqvist, until Jan. 2007
Jaakko Luttinen
Ilari Nieminen
Kristian Nybo
Markus Ojala
Juuso Parkkinen
Veli Peltola
Antti Puurula
Antti Rasinen
Rami Rautkorpi, until Jun. 2006
Ulrike Scharfenberger, until Apr. 2007
Jan-Hendrik Schleimer, until Aug. 2007
Santeri Seppä, until Feb. 2006
Jaakko Talonen
Seppo Virtanen
Paul Wagner
Aleksi Wallenius
Qi Yu, until Oct. 2007

Support staff

Leila Koivisto, department secretary
Tarja Pihamaa, laboratory secretary
Markku Ranta, B.Eng., works engineer
Miki Sirola, D.Sc. (Tech.), laboratory engineer
Mika Kongas, maintenance assistant, until Oct. 2006
Tapio Leipälä, maintenance assistant
Kimmo Rantala, maintenance assistant, until Sept. 2007
Petteri Räisänen, maintenance assistant, until Oct. 2007

Awards and activities

Prizes and scientific honours received by researchers of the unit

Professor Erkki Oja:

- IEEE Computational Intelligence Society Pioneer Award 2006, USA.

Professor Samuel Kaski:

- Distinguished Contribution Award, The 5th International Workshop on Mining and Learning with Graphs, Italy, 2007.

Dr. Alexander Ilin:

- Best student presentation at Fifth Conference on Artificial Intelligence Applications to Environmental Science as part of the 87th Annual Meeting of the American Meteorological Society, USA, 2007.

Dr. Mikko Kurimo:

- Professeur Inviteé 2005-2006 at Université de Saint-Etienne, France, 2006.

Dr. Jaakko Peltonen:

- Dr's Thesis Award 2004-2005, Pattern Recognition Society of Finland, 2006.

Important international positions of trust held by researchers of the unit

Professor Erkki Oja:

- Steering Committee Member and Session Chairman, Workshop on ICA and BSS, Charleston, USA, March 6-8, 2006.
- Program Committee Member, 2006 International Joint Conference on Neural Networks, IJCNN, Vancouver, Canada, July 17-20, 2006.
- Plenary talk "Emergence of semantics from multimedia databases," 2006 International Joint Conference on Neural Networks, IJCNN, Vancouver, Canada, July 17-20, 2006.
- Plenary talk "Emergence of semantics from multimedia databases," Third International Symposium on Neural Networks, ISNN, Chengdu, China, May 28-31, 2006.
- Plenary talk "Finding hidden factors in large spatiotemporal data sets," Machine Learning for Signal Processing XVII, Thessaloniki, Greece, Aug. 27-29, 2007.
- Plenary talk "Semantics, self-organizing maps, and multimedia databases," 7th Int. Conf. on Intelligent Systems Design and Applications (ISDA07), Rio de Janeiro, Brazil, Nov. 22-24. 2007.
- Panelist on the subject "challenges and opportunities of neural network research," Third International Symposium on Neural Networks, ISNN, Chengdu, China, May 28-31, 2006.
- Governing Board Member, International Neural Network Society, INNS, USA. Chairman of the Awards Committee.
- Member of the Executive Committee, Past President, European Neural Network Society, ENNS, The Netherlands, 2007.
- Member of the CELEST Advisory Board, Boston, USA.
- Member of the panel evaluation, NSF, Washington, DC, USA, May 22, 2006.
- Editorial Board Member:
  Natural Computing - An International Journal, The Netherlands.
  Neural Computation, USA.
  International Journal of Pattern Recognition and Artificial Intelligence, Singapore.
- Editor: Kollias, S., Stafylopatis, A., Duch, W., and Oja, E. (Eds): *Artificial Neural Networks - ICANN 2006. Volume I. Lecture Notes in Computer Science 4131.* Berlin, Germany: Springer (2006).
- Editor: Kollias, S., Stafylopatis, A., Duch, W., and Oja, E. (Eds): *Artificial Neural Networks - ICANN 2006. Volume II. Lecture Notes in Computer Science 4132.* Berlin, Germany: Springer (2006).
- Opponent at the doctoral dissertation of Michael Syskind Pedersen, DTU, Denmark, 2006.
- Opponent at the doctoral dissertation of Frederic Vrins, Université de Louvain, Belgium, 2007.

Professor Juha Karhunen:

- Program Committee Member:
  6th Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA2006), Charleston, South Carolina, USA, March 5-8, 2006.
  14th European Symposium on Artificial Neural Networks (ESANN2006), Bruges, Belgium, April 2006.
  7th Int. Conf. on Intelligent Data Engineering and Automated Learning (IDEAL2006), Burgos, Spain, Sept. 2006.
  The 15th European Symposium on Artificial Neural Networks (ESANN2007), Bruges, Belgium, April 2007.
  The 2007 IEEE International Joint Conference on Neural Networks (IJCNN2007), Orlando, Florida, USA, August 2007 USA.
  IEEE 2007 Workshop on Machine Learning in Signal Processing (MLSP2007), Thessaloniki, Greece, August 2007.
  The 7th International Conference on Independent Component Analysis and Blind Source Separation (ICA 2007), London, UK, September 2007.
  The 2nd Int. Workshop on Hybrid Artificial Intelligence Systems (HAIS'07), Salamanca, Spain, November 2007.
  The 8th Int. Conf. on Intelligent Data Engineering and Automated Learning (IDEAL'07), Birmingham, UK, December 2007.

- Session Chairman, 2006 Int. Conf. on Artificial Neural Networks (ICANN2006), Athens, Greece, Sept. 10-13, 2006.

- Evaluator in filling the academic chair of professor in bioinformatics, UMIT (Univ. for Health Sciences, Medical Informatics and Technology), Austria, 2006.

Professor Samuel Kaski:

- Invited talk "Implicit feedback from eye movements for proactive information retrieval," NIPS 2006 Workshop on User Adaptive Systems, Whistler, Canada, Dec. 8, 2006.

- Invited talk "Exploratory fusion of

  high-throughput data," 9th Northern European Bioinformatics conference–Bioinformatics 2007, Umea, Sweden, June 4-7, 2007.

- Invited talk "Machine learning for exploratory fusion and visualization of high-throughput data," Symposium on Bioinformatics and Chemical Genomics,

  Kyoto University, Japan, Sept. 20-21, 2007.

- Invited talk "Modeling of relevance," Brain-Inspired Information Technology

  (BrainIT2007), Hibikino, Kitakyushu, Japan, Nov. 13-16, 2007.

- Visiting lectures:
  "Proactive information retrieval," Intelligent sound seminar, Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark, Jan. 23, 2006.
  "Discriminative and associative clustering, and

  data fusion," Dagstuhl seminar 07131, "Similarity-based Clustering

and its Application to Medicine and Biology," Dagstuhl, Germany, March 30, 2007.
"Machine learning for data fusion and information visualization" in the invitation-only BSB 2007, First Bertinoro Systems Biology Workshop, Bertinoro, Italy, May 24, 2007.

"Relevance in data fusion and visualization," Bioinformatics Open Day, Chalmers University of Technology, Göteborg, Sweden, Dec. 10, 2007.

- Program Committee Member:
  International Symposium on Intelligent Data Engineering and Automated Learning, IDEAL'06, Burgos, Spain, 2006.
  ECML/PKDD'06, European Conference on Machine Learning and European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, 2006.
  New Trends in Intelligent Information Processing and Web Mining, IIPWM'06, Zakopane, Poland, June 19-22, 2006.
  Atlantic Web Intelligence Conference, AWIC'06, Beer-Sheva, Israel, 2006.
  IEEE International Workshop on Machine Learning for Signal Processing, MLSP'06, Maynooth, Ireland, Sept. 6-8, 2006.
  IEEE/WIC International Conference on Web Intelligence, WI 2006, Hong Kong, Dec. 18-22, 2006.
  International Conference on Natural Computation ICNC'06, Xi'an, China, Sept. 24-27, 2006.
  International Workshop on Web Semantics, WebS 2006, Krakow, Poland, Sept. 4-8, 2006.
  International Symposium on Neural Networks (ISNN2006), Chengdu, China, 2006.
  IIiX, the First Symposium on Information Interaction in Context, Copenhagen, Denmark, Oct. 18-20, 2006.
  MLMI-06, 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Washington DC, USA, May 1-3, 2006.
  SCAI 2006, The Ninth Scandinavian Conference on Artificial Intelligence, Espoo, Finland, Oct. 25-27, 2006.
  International Symposium on Intelligent Data Engineering and Automated Learning, IDEAL'07, Birmingham, UK, Dec. 16-19, 2007.
  Workshop on Self-Organizing Maps, WSOM'07, Bielefeld, Germany, Sept. 3-6, 2007.
  European Conference on Machine Learning and European Conference on Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD'07, Warsaw, Poland, Sept. 17-21, 2007.
  Atlantic Web Intelligence Conference AWIC'07, Fontainebleau, France, June 25-27, 2007.
  IEEE International Workshop on Machine Learning for Signal Processing, Thessaloniki, Greece, Aug. 27-29, 2007.
  IEEE/WIC International Conference on Web Intelligence, Silicon Valley, USA, Nov. 2-5, 2007.
  Multimodal Interaction and Related Machine Learning Algorithms MLMI'07, Brno, Czech Republic, June 28-30, 2007.
  European Symposium on Artificial Neural Networks ESANN'2007, Bruges, Belgium,

April 25-27, 2007.
The Fourth International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2007), Portofino, Italy, Oct. 7-10, 2007.
Machine Learning for Structural and Systems Biology, Evry, France, June 28-29, 2007.
ICML 2007, International Conference on Machine Learning, Corvallis, OR, USA, June 20-24, 2007.
ICCN'07, International Conference on Cognitive Neurodynamics, Shanghai, China, Nov. 17-21, 2007.
Data Mining in Functional Genomics and Proteomics: Current Trends and Future Directions, Workshop in ECML-PKDD 2007, Warsaw, Poland, Sept. 17, 2007.

- Session Chairman, 9th Northern European Bioinformatics Conference, Bioinformatics 2007, Umea, Sweden, June 4-7, 2007.

- Editorial Board Member:
Intelligent Data Analysis, The Netherlands
International Journal of Neural Systems, Singapore
Cognitive Neurodynamics, Germany.

- Evaluator in filling the academic chair of professor of bioinformatics, NUI Galway, Ireland, 2006.

- Evaluator in filling the academic chair of professor of computer science, University of Southampton, UK, 2007.

- Opponent at the doctoral dissertation of Rasmus Elborg Madsen, Technical University of Denmark, 2006.

Academy Professor Heikki Mannila:

- Program Committee Member:
Twelth SIGKDD Conference on Data Mining and Knowledge Discovery (KDD'06), Philadelphia, USA, Aug. 20-23, 2006.
ACM SIGMOD Symposium on Management of Data (SIGMOD) 2006, Chicago, USA, June 26-29, 2006.
Combinatorial Pattern Matching (CPM 2007), London, Canada, July 9-11, 2007.

- Best paper chair, 13th ACM SIGKDD Conference on Data Mining and Knowledge Discovery (KDD'07), San Jose, CA, USA, Aug. 12-15, 2007.

- Member of the ESFRI Physical Sciences and Engineering Roadmap Working Group, Luxembourg.

- Chairman of the ESFRI Expert Group on Computation and Data Treatment, Luxembourg.

- Member of the evaluation panel of Computer Science at Uppsala University, Sweden, 2007.

- Area Editor, IEEE Transactions Knowledge and Data Engineering, USA.

- Associate Editor, Transactions on Database Systems, USA.

- Editorial Board Member, ACM Transactions on Knowledeg Discovery in Data, USA.

- Member of Steering Committee, Data Mining and Knowledge Discovery, USA.
- Opponent at the doctoral dissertation of Niklas Noren, Stockholm University, 2007.

Professor Olli Simula:

- Session Chairman and Program Committee Member, International Conference on Artificial Neural Networks, ICANN 2006, Athens, Greece, Sept. 10-14, 2006.
- Program Committee Member:
  2006 IEEE Mountain Workshop on Adaptive and Learning Systems, SMCals/06, Logan, Utah, USA, July 24-26, 2006.
  International Conference on Advances in Pattern Recognition, Kolkata, India, Jan. 2-4, 2007.
  6th International Workshop on Self-Organizing Maps, WSOM'07, Bielefeld, Germany, Sept. 3-6, 2007.
  International Conference on Artificial Neural Networks, ICANN 2007, Porto, Portugal, Sept. 9-13, 2007.
- Executive Committee Member, European Neural Network Society (ENNS), The Netherlands.
- Scientific Council Member, Institute Eurecom, France.

Dr. Gemma Garriga:

- Editor of the Special Issue on Mining and Learning in Graphs at *Machine Learning Journal*, USA, 2007.

Dr. Jaakko Hollmén:

- Referee for the National Fund for Scientifific & Technological Development (FONDECYT), Chile, 2007.
- Program Committee Member:
  The 17th European Conference on Machine Learning (ECML 2006) and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006), Berlin, Germany, 2006.
  7th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'06), Burgos, Spain, 2006.
  8th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2006), Krakow, Poland, 2006.
  The 18th European Conference on Machine Learning (ECML 2007) and the 11th European Conference on Principles and Practice of Knowledge Discovery in databases (PKDD 2007), Warsaw, Poland, 2007.
  The 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2007), Birmingham, United Kingdom, 2007.
  9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007), Regensburg, Germany, 2007.

Dr. Antti Honkela:

- Program Committee Member, The Second Annual Conference on Bio-Inspired Computing: Theory and Applications (BIC-TA 2007), Zheng Zhou, China, Sept. 14-17, 2007.

- Visiting lectures:
  "Variational Bayesian learning in nonlinear BSS and state-space modelling", Technical University of Denmark, Informatics and Mathematical Modelling, April 5, 2006.
  "Introduction to expectation propagation", Technical University of Denmark, Informatics and Mathematical Modelling, August 16, 2006.
  "Modelling nonlinear dynamical systems by variational Bayesian learning," University of Glasgow, Dept. of Computing Science, U.K., Nov. 7, 2007.
  "Modelling nonlinear dynamical systems by variational Bayesian learning," University of Manchester, School of Computer Science, U.K., Nov. 14, 2007.
  "A tutorial on expectation propagation," University of Manchester, School of Computer Science, Nov. 20, 2007.

Dr. Timo Honkela:

- Visiting lectures:
  "Learning to translate,"

  Scuola Normale Superiore, Pisa, Italy,

  April 28, 2006.
  "Issues in knowledge representation and reasoning,"

  National Center for Scientific Research "Demokritos", Athens, Greece

  June 6, 2006.
  "Social artificial intelligence:

  Mobile devices that learn to read between the lines,"

  Nokia Research Center Palo Alto, Palo Alto, California, USA,

  Nov. 15, 2006.

  "Multidisciplinary view of consumer models:

  Hobbyism and playful strategies meet adaptive models,"

  Wallenberg Hall, Stanford University, California, USA,

  Nov. 17, 2006 (with Tanja Kotro and Sari Stenfors).

  "Inherent fuzziness of language:

  Cognitive, philosophical and computational aspects,"

  Berkeley Initiative on Soft Computing

  University of California Berkeley, Berkeley, USA,

  Nov, 20, 2006.

  "Language, learning and multimodal systems: Technological,

  cognitive and philosophical perspectives,"

  University of California Santa Barbara, Santa Barbara, USA,

  Dec. 1, 2006.

- Chair of the Program Committee and Session Chairman, Scandinavian Conference on Artificial Intelligence, Espoo, Finland, Oct. 25-27, 2006.

- Program Committee Member:
  Artificial Intelligence Applications and Innovations (AIAI) 2006, Athens, Greece, Sept. 7-9, 2006.
  2nd International Nonlinear Science Conference Heraklion, Crete, Greece, March 10-12, 2006.

- Session Chairman:
  Third International Symposium on the Emergence and Evolution of Linguistic Communication, EELC III, Rome, Italy, Sept. 30 - Oct. 1, 2006.
  International Joint Conference on Neural Networks 2007, Orlando, Florida, USA, Aug. 12-17, 2007.

- Chairman of the Working Group 12.1 (Knowledge Representation and Reasoning), IFIP (International Federation for Information Processing), Austria.

- Editorial Board Member, Constructivist Foundations, Austria.

- Referee of EU project proposals, 2007.

- Opponent at the doctoral dissertation of Toomas Kirt, Tallinn University of Technology, Estonia, 2007.

Dr. Markus Koskela:

- Program Committee Member:
  9th International Conference series on Visual Information Systems, Shanghai, China, June 28-29, 2007.
  Pacific-Rim Conference on Multimedia (PCM), Hong Kong, China, Dec. 11-14, 2007.

Dr. Mikko Kurimo:

- Invited talk "Unsupervised segmentation of words into morphemes – Challenge 2005: An Introduction and Evaluation Report," PASCAL Challenge Workshop 2006, Venice, Italy, April 12, 2006.

- Visiting lecture "Unlimited vocabulary language models," Université de Saint-Etienne, France, May 2006.

- Chairman of the Program Committee, Unsupervised segmentation of words into morphemes, PASCAL Challenge Workshop 2006, Venice, Italy, April 12, 2006.

- Program Committee Member, Morpho Challenge Workshop, Budapest, Hungary, Sept. 19, 2007.

- Editor of the Proceedings of the PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes, Venice, Italy, 2006 (PASCAL European Network of Excellence, Italy).

- Opponent at the doctoral dissertation of Tanel Alumäe, Tallinn University of Technology, Estonia, 2006.

Dr. Jorma Laaksonen:

- Program Committee Member, 9th International Conference series on Visual Information Systems, Shanghai, China, June 28-29, 2007.

Dr. Amaury Lendasse:

- Member of the PhD jury of Luis Javier Herrera, University of Granada, Spain, July 5, 2007.
- Member of the PhD jury of Alberto Guillen, University of Granada, Spain, July 6, 2007.
- Seminar on the subject "Variable selection," University of Granada, Spain, July 9, 2007.

Dr. Jaakko Peltonen:

- Program Committee Member, 9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007), Regensburg, Germany, Sept. 3-9, 2007.

Dr. Tapani Raiko:

- Organizing committee chair and program committee member, Scandinavian Conference on Artificial Intelligence, Espoo, Finland, Oct. 25-27, 2006.
- Program committee member, European Conference on Machine Learning (ECML 2007), Warsaw, Poland, Sept. 17-21, 2007.

Dr. Jouni Seppänen:

- Program Committee Member:
  ECML/PKDD 2007, Warsaw, Poland, Sept. 17-21, 2007.
  ACM SAC (Symposium on Applied Computing), data mining track, Seoul, Korea, March 11-15, 2007.

Dr. Miki Sirola:

- Program Committee Member:
  International Conference on Modelling, Identification and Control, Lanzarote, Spain, Feb. 6-8, 2006.
  International Conference on Applied Simulation and Modelling, Rhodes, Greece, June 26-28, 2006.
  International Conference on Modelling and Simulation, Montreal, Canada, May 24-26, 2006.
  International Conference on Intelligent Systems and Control, Honolulu, USA, August 14-16, 2006.
  International Conference on Modelling, Identification and Control Innsbruck, Austria, Feb. 12-14, 2007.
  International Conference on Modelling and Simulation, Montreal, Canada, May 30 - June 1, 2007.
  International Conference on Computational Intelligence, Banff, Canada, July 2-4, 2007.
  IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2007), Dortmund, Germany, Sept. 6-8, 2007.
  International Conference on Intelligent Systems and Control, Cambridge, USA, Nov. 19-21, 2007.

M.Sc. Mika Sulkava:

- Invited talk "Analysis of carbon flux data," Mini-Workshop on Interannual Variability of CO_2 flux of forests, Antwerp, Belgium, Mar. 24, 2006.

Important domestic positions of trust held by researchers of the unit

Professor Erkki Oja:

- Chairman, Research Council for Natural Sciences and Engineering, Academy of Finland, 2007.
- Vice Chairman, Finnish Academy of Sciences and Letters, Group of Mathematics and Computer Science.
- Vice Chairman, Alfred Kordelin Foundation.
- Chair, IEEE Computational Intelligence Chapter, Finland.
- Opponent at the doctoral dissertation of Leena Lepistö, Tampere University of Technology, 2006.
- Opponent at the doctoral dissertation of Olga Kayo, University of Oulu, 2006.
- Opponent at the doctoral dissertation of Markus Turtinen, University of Oulu, 2007.
- Opponent at the doctoral dissertation of Mari Partio, Tampere University of Technology, 2007.
- Opponent at the doctoral dissertation of Hannu Laamanen, University of Joensuu, 2007.

Professor Juha Karhunen:

- Member of Scientific Committee, European Symposium on Time Series Prediction, ESTSP'07, Espoo, Finland, Feb. 7-9, 2007.

Professor Samuel Kaski:

- Visiting lecture "Supervised data mining and dependency exploration for bioinformatics," BioCity Turku, Jan. 17, 2006.
- Visiting talks:
  "Adaptiivinen informatiikka," VTT research seminar, Espoo, Apr. 12, 2006.
  "Adaptive informatics," Nokia-TKK Research Seminar, Nokia Research Center, Helsinki, May 23, 2006.
  "Search without typing - proactive retrieval system anticipates your interests," Publication seminar of the Research Programme on Proactive Computing Evaluation Report, Academy of Finland, Helsinki, Feb. 28, 2007.

  "Probabilistic adaptive systems," HIIT-NRC research seminar, Nokia Research Center, Helsinki, Oct. 3, 2007.
- Member of thesis committee, Pattern Recognition Society of Finland, 2007.
- Opponent at the doctoral dissertation of Jussi Salmi, University of Oulu, 2006.

- Evaluator in filling the academic chair of professor of distributed systems, University of Jyväskylä, 2006.
- Evaluator in filling the academic chair of professor of computer science, University of Helsinki, 2006.
- Referee of docentships on computer science (2) and bioinformatics, University of Helsinki, 2006.
- Participation in the FinnSight 2015 foresight project of the Academy of Finland and the National Technology Agency Tekes, panel on bio-knowledge and society, 2006.
- Program Committee Member:
  Probabilistic Modeling and Machine Learning in Structural and Systems Biology, Tuusula, Finland, June 17-18, 2006.
  IEEE International Workshop on Genomic Signal Processing and Statistics, GENSIPS'07, Tuusula, Finland, June 10-12, 2007.
- Member of Scientific Committee, European Symposium on Time Series Prediction, ESTSP'07, Espoo, Finland, Feb. 7-9, 2007.

Academy Professor Heikki Mannila:

- Opponent at the doctoral dissertation of Sami Äyrämö, University of Jyväskylä, 2006.
- Evaluator in filling the academic chair of professor of interactive technology, University of Tampere, 2006.
- Evaluator in filling the academic chair of professor of computer science, University of Kuopio, 2006.

Professor Olli Simula:

- Chairman, IEEE Computer Chapter, Finland.
- Opponent at the doctoral dissertation of Ulla Elsilä, University of Oulu, 2007.
- Member of Scientific Committee, European Symposium on Time Series Prediction, ESTSP'07, Espoo, Finland, Feb. 7-9, 2007.
- Program Committee Member, 1st IEEE WoWMoM Workshop on Autonomic Wireless Access, IWAS07, Helsinki, Finland, June 18th, 2007.

Dr. Jaakko Hollmén:

- Referee for The Finnish Work Environment Fund, Finland, 2007.
- Opponent at the doctoral dissertation of Miika Ahdesmäki, Tampere University of Technology, Department of Information Technology, 2007.
- Invited talks:
  "Data analysis in environmental informatics," Finnish Forest Research Center, Vantaa, Finland May 14, 2007.
  "Modeling DNA copy number amplification patterns in human cancers," University

of Helsinki, Department of Computer Science, Feb. 15, 2007.
"Modeling DNA copy number amplification patterns in human cancers," GeneOS Ltd., Helsinki, May 7, 2007.
"Modeling DNA copy number amplification patterns in human cancers," Turku Centre for Biotechnology, Dec. 11, 2007.

- Member of Scientific Committee, European Symposium on Time Series Prediction, ESTSP'07, Espoo, Finland, Feb. 7-9, 2007.

Dr. Timo Honkela:

- Seminar talk "Tulevaisuuden kuluttaja,"
  seminar of the project "KULTA", Vierumäki, Aug.
  15, 2006.
- Visiting talk "Social artificial intelligence:
  Mobile devices that learn to read between the lines,"
  Nokia Research Center, Helsinki,
  Oct. 11, 2006.
- Editorial Board Member, Puhe ja kieli, Finland.

Dr. Markus Koskela:

- Governing Board Member, Pattern Recognition Society of Finland, 2007.

Dr. Mikko Kurimo:

- Member of the Executive Committee, IEEE Finland Section, 2006.
- Treasurer, IEEE Finland Section.
- Opponent at the doctoral dissertation of Marko Moberg, Tampere University of Technology, 2007.

Dr. Jorma Laaksonen:

- Chairman of the dictionary committee, Pattern Recognition Society of Finland.

Dr. Amaury Lendasse:

- Organizing Committee Member, European Symposium on Time Series Prediction, ESTSP'07, Espoo, Finland, Feb. 7-9, 2007.

M.Sc. Elia Liitiäinen:

- Organizing Committee Member, European Symposium on Time Series Prediction, ESTSP'07, Espoo, Finland, Feb. 7-9, 2007.

Dr. Tapani Raiko:

- Chairman of the Finnish Artificial Intelligence Society, 2006-2007.
- Chair of the organizing committee, Finnish Artificial Intelligence Conference, Espoo, Finland, Oct. 26-27, 2006.

M.Sc. Antti Sorjamaa:

- Organizing Committee Member, European Symposium on Time Series Prediction, ESTSP'07, Espoo, Finland, Feb. 7-9, 2007.

Research visits abroad by researchers of the unit; 2 weeks or more

- Dr. Mathias Creutz, International Computer Science Institute (ICSI), Berkeley, CA, USA, Jan.-Jun. 2006 (6 mths).
- Dr. Antti Honkela, Technical University of Denmark, Mar.-May 2006 (2 mths).
- Dr. Timo Honkela, Stanford University, CA, USA, Nov. 2006 (2 wks).
- Dr. Ville Könönen, Vrije Universiteit Brussel, Belgium, Apr.-July 2006 (3 mths).
- Dr. Markus Koskela, Dublin City University, Centre for Digital Video Processing, Ireland, Jan. 2006-Jan. 2007 (13 mths).
- Dr. Mikko Kurimo, Université de Saint-Etienne, France, May 2006 (1 mth).
- Dr. Janne Nikkilä, European Bioinformatics Institute, Hinxton, U.K., Nov.-Dec. 2006 (2 mths).
- M.Sc. Hannes Heikinheimo, Google Switzerland, Zürich, May-Aug. 2007 (3 mths).
- Dr. Antti Honkela, University of Manchester, School of Computer Science, U.K., Sep.-Nov. 2007 (3 mths).
- Professor Samuel Kaski, European Bioinformatics Institute EBI, Hinxton, U.K., Oct.-Dec. 2007 (2 mths).
- Dr. Mikko Kurimo, Stanford Research Institute, SRI, Nov. 2007- (1 mth).
- M.Sc. Elia Liitiäinen, Université catholique de Louvain, Louvain-la-Neuve, Belgium, Oct.-Nov. 2007 (2 mths).
- Dr. Jussi Pakkanen, University of Melbourne, Australia, Feb.-Aug. 2007 (6 mths).
- Dr. Kalle Palomäki, The University of Sheffield, U.K., Jan-Feb. 2007, Dec. 2007 (3 wks).
- Dr. Ricardo Vigário, Grenoble Institute of Technology, France, Nov. 2007 - (2 mths).

Research visits by foreign researchers to the unit; 2 weeks or more

- M.Sc. Ane Amaia Orue-Etxebarria Apellaniz, University of the Basque Country, Spain, Nov. 2006–.
- Dr. Jacob Golderger, Bar-Ilan University, Israel, Aug.-Sep. 2006 (3 wks).
- Dr. David Hardoon, University of Southampton, U.K., Mar.& Nov. 2006 (2 wks).
- Dr. Robert Gwadera, University of Purdue, USA, –Jul. 2007.
- Dr. Gemma Garriga, Universitat Politecnica de Catalunya, Barcelona, Spain, Oct. 2006–.
- B.Sc. Jan-Hendrik Schleimer, University of Tübingen, Germany, –Aug. 2007.

- Dr. Patrik Bas, LIS, INPG, France, –Dec. 2007.
- Prof. Aurelio Campilho, University of Porto, Biomedical Engineering Institute, Portugal, Jun.-Aug. 2006 (2 mths).
- M.Sc. Ester Gonzalez, UAM (Universidad Autonoma de Madrid), Spain, Jul.-Sep. 2006.
- Prof. Nobuo Matsuda, Oshima College of Maritime Technology, Information Science & Technology Dept., Japan, –Mar. 2006.
- Dr. César Fernández, Miguel Hernández University, Spain, Aug.-Nov. 2006.
- Dr. Francesco Corona, Universita degli Studi di Cagliari, Dipartimento di Ingegneria Chimica e Materiali, Italy, Sept.-Oct. 2006, Jan. 2007–.
- M.Sc. Yoan Miché, Institut National Polytechnique de Grenoble, France, Feb. 2006-.
- Mark van Heeswijk, Technical University of Eindhoven, The Netherlands, May 2007-.
- M.Sc. Jose Vellez Caldas, INESC, Portugal, Aug. 2007–.
- M.Sc. Yusuf Yaslan, Istanbul Technical University, Turkey, Sep. 2007–.
- M.Sc. Maarten Peeters, Vrije Universiteit, Brussel, Belgium, Jan.-Mar. 2007 (2 mths).
- Dr. Franck Thollard, Université Jean Monnet, Saint-Étienne, France, May 2007 (4 wks).
- M.Sc. Philip Prentis, Czech Technical University, Prague, Czech Republic, May-Sept. 2007.
- Federico Montesino Pouzols, University of Seville, Spain, Jun.-Aug. 2007.
- B.Sc. Ulrike Scharfenberger, University of Tübingen, Germany, Sep. 2006- Apr. 2007.
- David Ellis, Brown University, Providence, RI, USA, Mar.-Jun. 2007 (3 mths).
- Paul Grouchy, Queen's University, Kingston, Canada, Jun.-Aug. 2007.
- M.Sc. Fernando Mateo Jimenez, Universidad Politecnica de Valencia, Spain, Sep.-Nov. 2007.
- M.Sc. Marcin Blachnik, Silesian University of Technology, Katowice, Poland, Oct. 2007–.

Other activities

Dr. Mikko Kurimo:

- interview on subject "Status of speech recognition research" for Radio YleX Studio, Aug. 30, 2006.
- interview on subject "Speech recognition is a great innovation," for the journal T-lehti No. 5, 2006 (in Finnish).

Dr. Tapani Raiko:

- interview on the subject "Artificial intelligence" for the weekly magazine Valo of Aamulehti, 2006.

Courses

Courses given by the Laboratory of Computer and Information Science.

Spring Semester 2006

Code	Course	Lecturer	Course Assistant
T-61.124	Special Project in Computer Architecture	S. Haltsonen	A. Sorjamaa
T-61.152	Seminar on Computer and Information Science	A. Honkela	
T-61.3010	Digital Signal Processing and Filtering	O. Simula	J. Parviainen, T. Similä, A. Sorjamaa
T-61.3020	Principles of Pattern Recognition	E. Oja	T. Raiko, M. Aksela
T-61.3030	Principles of Neural Computing	K. Raivio	J. Venna, M. Aksela
T-61.5010	Information Visualization	K. Puolamäki	A. Ukkonen
T-61.5020	Statistical Natural Language Processing	T. Honkela, T. Hirsimäki	S. Virpioja
T-61.5040	Learning Models and Methods	P. Pajunen	V. Viitaniemi
T-61.5070	Computer Vision	J. Laaksonen	R. Rautkorpi
T-61.5090	Image Analysis in Neuroinformatics	R. Vigário, J. Särelä	
T-61.6020	Special Course II: *Reinforcement Learning – Theory and Applications*	V. Könönen	
T-61.6030	Special Course III: *Independent Component Analysis*	J. Karhunen	M. Harva
T-61.6060	Special Course VI: *Data Analysis and Environmental Informatics*	J. Hollmén, M. Sulkava	H. Heikinheimo
T-61.6070	Special Course in Bioinformatics I: *Computational Modeling Methods in Bioinformatics*	S. Kaski	M. Oja

Fall Semester 2006

Code	Course	Lecturer	Course Assistant
T-61.2010	From Data to Knowledge	E. Oja, H. Mannila, J. Seppänen	U. Remes, J. Parviainen
T-61.3040	Statistical Signal Modeling	P. Pajunen	V. Viitaniemi
T-61.5030	Advanced Course in Neural Computing	J. Karhunen	J. Peltonen
T-61.5080	Signal Processing in Neuroinformatics	R. Vigário, J. Särelä	
T-61.5100	Digital Image Processing	J. Laaksonen	M. Sjöberg
T-61.6010	Special Course I: *Gaussian Processes for Machine Learning*	A. Honkela	S. Hanhijärvi
T-61.6040	Special Course IV: *Variable Selection for Regression*	A. Lendasse	E. Liitiäinen
T-61.6050	Special Course V: *Data Mining in Telecommunication*	K. Raivio	M. Multanen
T-61.6080	Special Course in Bioinformatics II: *Data Integration and Fusion in Bioinformatics*	J. Nikkilä	L. Lahti
T-61.6090	Special Course in Language Technology: *Finnish-Swedish Machine Translation Challenge*	T. Honkela M. Creutz	
T-61.6900	Individual Studies: *Reading Circle on Bayesian Theory*	S. Kaski	

Spring Semester 2007

Code	Course	Lecturer	Course Assistant
T-61.152	Seminar on Computer and Information Science	A. Honkela	
T-61.2020	From Data to Knowledge, Exercise Project		J. Parviainen
T-61.3010	Digital Signal Processing and Filtering	O. Simula	J. Parviainen, N. Reyhani T. Similä
T-61.3020	Principles of Pattern Recognition	E. Oja	T. Raiko, M. Aksela
T-61.3030	Principles of Neural Computing	K. Raivio, M. Koskela	M. Pöllä, M. Aksela
T-61.5010	Information Visualization	K. Puolamäki	A. Ukkonen
T-61.5020	Statistical Natural Language Processing	T. Honkela, T. Hirsimäki, M. Creutz	S. Virpioja
T-61.5040	Learning Models and Methods	P. Pajunen	V. Viitaniemi
T-61.5050	High-Throughput Bioinformatics	J. Nikkilä	M. Oja
T-61.5070	Computer Vision	J. Laaksonen	M. Sjöberg
T-61.5090	Image Analysis in Neuroinformatics	R. Vigário, J. Särelä	
T-61.6020	Special Course II: *Machine Learning: Basic Principles*	K. Puolamäki	M. Korpela
T-61.6030	Special Course III: *Introductory Elements of Functional Data Analysis*	F. Corona A. Lendasse	E. Liitiäinen
T-61.6060	Special Course VI: *Decision Support with Data Analysis*	M. Sirola	G. Lampi
T-61.6070	Special Course in Bioinformatics I: *Modeling of Biological Networks*	S. Kaski	A. Klami

Fall Semester 2007

Code	Course	Lecturer	Course Assistant
T-61.2010	From Data to Knowledge	E. Oja, H. Mannila, J. Seppänen	U. Remes, V. Viitaniemi
T-61.3040	Statistical Signal Modeling	J. Peltonen	V. Viitaniemi
T-61.3050	Machine Learning: Basic Principles	K. Puolamäki	A. Ukkonen
T-61.5060	Algorithmic Methods of Data Mining	H. Mannila, K. Puolamäki	N. Vuokko
T-61.5080	Signal Processing in Neuroinformatics	R. Vigário	J. Ylipaavalniemi
T-61.5100	Digital Image Processing	J. Laaksonen	M. Sjöberg
T-61.5110	Modeling Biological Networks	S. Kaski, P. Auvinen	A. Klami
T-61.5120	Computational Genomics	S. Hautaniemi	K. Nousiainen
T-61.5130	Machine Learning and Neural Networks	J. Karhunen	M. Pöllä
T-61.6040	Special Course IV: *Information Networks*	G. C. Garriga	N. Tatti
T-61.6050	Special Course V: *Nonlinear Dimensionality Reduction*	A. Lendasse F. Corona	K. Nybo
T-61.6080	Special Course in Bioinformatics II: *Prior Knowledge and Background Data in Computational Inferenceproteomic and Metabolic Data*	J. Nikkilä	L. Lahti
T-61.6090	Special Course in Language Technology: *Emergence of Cognition, Communication and Language: from Humans to Machines*	K. Lagus T. Honkela	O. Kohonen

Doctoral dissertations

Using and Extending Itemsets in Data Mining: Query Approximation, Dense Itemsets, and Tiles

Jouni K. Seppänen

Dissertation for the degree of Doctor of Science in Technology on 31 May 2006.

External examiners:
Gautam Das (University of Texas at Arlington)
Bart Goethals (University of Antwerp)
Opponent:
Dimitrios Gunopulos (University of California at Riverside)

Abstract:
Frequent itemsets are one of the best known concepts in data mining, and there is active research in itemset mining algorithms. An itemset is frequent in a database if its items co-occur in sufficiently many records. This thesis addresses two questions related to frequent itemsets. The first question is raised by a method for approximating logical queries by an inclusion-exclusion sum truncated to the terms corresponding to the frequent itemsets: how good are the approximations thereby obtained? The answer is twofold: in theory, the worst-case bound for the algorithm is very large, and a construction is given that shows the bound to be tight; but in practice, the approximations tend to be much closer to the correct answer than in the worst case. While some other algorithms based on frequent itemsets yield even better approximations, they are not as widely applicable.

The second question concerns extending the definition of frequent itemsets to relax the requirement of perfect co-occurrence: highly correlated items may form an interesting set, even if they never co-occur in a single record. The problem is to formalize this idea in a way that still admits efficient mining algorithms. Two different approaches are used. First, dense itemsets are defined in a manner similar to the usual frequent itemsets and can be found using a modification of the original itemset mining algorithm. Second, tiles are defined in a different way so as to form a model for the whole data, unlike frequent and dense itemsets. A heuristic algorithm based on spectral properties of the data is given and some of its properties are explored.

Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition

Mathias Creutz

Dissertation for the degree of Doctor of Science in Technology on 15 June 2006.

External examiners:
Richard Wicentowski (Swarthmore College, Pennsylvania)
Jukka Heikkonen (Helsinki University of Technology)
Opponents:
James H. Martin (University of Colorado)
Wray Buntine (Helsinki Institute of Information Technology)

Abstract:
In order to develop computer applications that successfully process natural language data (text and speech), one needs good models of the vocabulary and grammar of as many languages as possible. According to standard linguistic theory, words consist of morphemes, which are the smallest individually meaningful elements in a language. Since an immense number of word forms can be constructed by combining a limited set of morphemes, the capability of understanding and producing new word forms depends on knowing which morphemes are involved (e.g., "water, water+s, water+y, water+less, water+less+ness, sea+water").

Morpheme boundaries are not normally marked in text unless they coincide with word boundaries. The main objective of this thesis is to devise a method that discovers the likely locations of the morpheme boundaries in words of any language. The method proposed, called Morfessor, learns a simple model of concatenative morphology (word forming) in an unsupervised manner from plain text. Morfessor is formulated as a Bayesian, probabilistic model. That is, it does not rely on predefined grammatical rules of the language, but makes use of statistical properties of the input text.

Morfessor situates itself between two types of existing unsupervised methods: morphology learning vs. word segmentation algorithms. In contrast to existing morphology learning algorithms, Morfessor can handle words consisting of a varying and possibly high number of morphemes. This is a requirement for coping with highly-inflecting and compounding languages, such as Finnish. In contrast to existing word segmentation methods, Morfessor learns a simple grammar that takes into account sequential dependencies, which improves the quality of the proposed segmentations.

Morfessor is evaluated in two complementary ways in this work: directly by comparing to linguistic reference morpheme segmentations of Finnish and English words and indirectly as a component of a large (or virtually unlimited) vocabulary Finnish speech recognition system. In both cases, Morfessor is shown to outperform state-of-the-art solutions.

The linguistic reference segmentations were produced as part of the current work, based on existing linguistic resources. This has resulted in a morphological gold standard, called Hutmegs, containing analyses of a large number of Finnish and English word forms.

Blind Source Separation for Interference Cancellation in CDMA Systems

Karthikesh Raju

Dissertation for the degree of Doctor of Science in Technology on 11 August 2006.

External examiners:
Lars Rasmussen (University of South Australia)
Asoke Nandi (University of Liverpool)
Opponent:
Jürgen Lindner (Universität Ulm)

Abstract:
Communication is the science of "reliable" transfer of information between two parties, in the sense that the information reaches the intended party with as few errors as possible. Modern wireless systems have many interfering sources that hinder reliable communication. The performance of receivers severely deteriorates in the presence of unknown or unaccounted interference. The goal of a receiver is then to combat these sources of interference in a robust manner while trying to optimize the trade-off between gain and computational complexity.

Conventional methods mitigate these sources of interference by taking into account all available information and at times seeking additional information e.g., channel characteristics, direction of arrival, etc. This usually costs bandwidth. This thesis examines the issue of developing mitigating algorithms that utilize as little as possible or no prior information about the nature of the interference. These methods are either semi-blind, in the former case, or blind in the latter case.

Blind source separation (BSS) involves solving a source separation problem with very little prior information. A popular framework for solving the BSS problem is independent component analysis (ICA). This thesis combines techniques of ICA with conventional signal detection to cancel out unaccounted sources of interference. Combining an ICA element to standard techniques enables a robust and computationally efficient structure. This thesis proposes switching techniques based on BSS/ICA effectively to combat interference. Additionally, a structure based on a generalized framework termed as denoising source separation (DSS) is presented. In cases where more information is known about the nature of interference, it is natural to incorporate this knowledge in the separation process, so finally this thesis looks at the issue of using some prior knowledge in these techniques. In the simple case, the advantage of using priors should at least lead to faster algorithms.

Approaches for Content-Based Retrieval of Surface Defect Images

Jussi Pakkanen

Dissertation for the degree of Doctor of Science in Technology on 20 October 2006.

External examiners:
Matti Niskanen (University of Oulu)
Andreas Rauber (Vienna University of Technology)
Opponent:
Pasi Koikkalainen (University of Jyväskylä)

Abstract:
There are two properties which all industrial manufacturing processes try to optimize: speed and quality. Speed can also be called throughput and tells how much products can be created in a specified time. The higher speeds you have the better. Quality means the perceived goodness of the finished product. Broken or defective products simply don't sell, so they must be eliminated.

These are contradicting goals. The larger the manufacturing volumes, the less time there is to inspect a single product, or the more inspectors are required. A good example is paper manufacturing. A single paper machine can produce a sheet of paper several meters wide and several hundred kilometers long in just a few hours. It is impossible to inspect these kinds of volumes by hand.

In this thesis the indexing and retrieval of defect images taken by an automated inspection machine is examined. Some of the images taken contain serious defects such as holes, while others are less grave. The goal is to try to develop automated methods to find the serious fault images from large databases using only the information in the images. This means that there are no annotations. This is called content-based image retrieval, or CBIR.

This problem is examined in two different ways. First the PicSOM CBIR tool's suitability for this task is evaluated. PicSOM is a platform for content-based image retrieval developed at the Laboratory of Computer and Information Science, Helsinki University of Technology. PicSOM has earlier been succesfully applied to various different CBIR tasks.

The other part involves developing new algorithms for efficient indexing of large, high-dimensional databases. The Evolving Tree (ETree), a novel hierarchical, tree-shaped, self-organizing neural network is presented and analyzed. It is noticeably faster than classical methods, while still obtaining good results.

The suitability and performance of both CBIR and ETree on this problem is evaluated using several different experiments. The results show that both approaches are applicable for this real world quality inspection problem with good results.

Advanced source separation methods with applications to spatio-temporal datasets

Alexander Ilin

Dissertation for the degree of Doctor of Science in Technology on 3 November 2006.

External examiners:
Mark A. Girolami, (University of Glasgow)
Aki Vehtari (Helsinki University of Technology)
Opponent:
Luis Borges de Almeida, (Technical University of Lisbon)

Abstract:

Latent variable models are useful tools for statistical data analysis in many applications. Examples of popular models include factor analysis, state-space models and independent component analysis. These types of models can be used for solving the source separation problem in which the latent variables should have a meaningful interpretation and represent the actual sources generating data. Source separation methods is the main focus of this work.

Bayesian statistical theory provides a principled way to learn latent variable models and therefore to solve the source separation problem. The first part of this work studies variational Bayesian methods and their application to different latent variable models. The properties of variational Bayesian methods are investigated both theoretically and experimentally using linear source separation models. A new nonlinear factor analysis model which restricts the generative mapping to the practically important case of post-nonlinear mixtures is presented. The variational Bayesian approach to learning nonlinear state-space models is studied as well. This method is applied to the practical problem of detecting changes in the dynamics of complex nonlinear processes.

The main drawback of Bayesian methods is their high computational burden. This complicates their use for exploratory data analysis in which observed data regularities often suggest what kind of models could be tried. Therefore, the second part of this work proposes several faster source separation algorithms implemented in a common algorithmic framework. The proposed approaches separate the sources by analyzing their spectral contents, decoupling their dynamic models or by optimizing their prominent variance structures. These algorithms are applied to spatio-temporal datasets containing global climate measurements from a long period of time.

Bayesian Inference in Nonlinear and Relational Latent Variable Models

Tapani Raiko

Dissertation for the degree of Doctor of Science in Technology on 1 December 2006.

External examiners:
Jouko Lampinen (Helsinki University of Technology)
Petri Myllymäki (University of Helsinki)
Opponent:
Ole Winther (Danmarks Tekniske Universitet)

Abstract:
Statistical data analysis is becoming more and more important when growing amounts of data are collected in various fields of life. Automated learning algorithms provide a way to discover relevant concepts and representations that can be further used in analysis and decision making.

Graphical models are an important subclass of statistical machine learning that have clear semantics and a sound theoretical foundation. A graphical model is a graph whose nodes represent random variables and edges define the dependency structure between them. Bayesian inference solves the probability distribution over unknown variables given the data. Graphical models are modular, that is, complex systems can be built by combining simple parts. Applying graphical models within the limits used in the 1980s is straightforward, but relaxing the strict assumptions is a challenging and an active field of research.

This thesis introduces, studies, and improves extensions of graphical models that can be roughly divided into two categories. The first category involves nonlinear models inspired by neural networks. Variational Bayesian learning is used to counter overfitting and computational complexity. A framework where efficient update rules are derived automatically for a model structure given by the user, is introduced. Compared to similar existing systems, it provides new functionality such as nonlinearities and variance modelling. Variational Bayesian methods are applied to reconstructing corrupted data and to controlling a dynamic system. A new algorithm is developed for efficient and reliable inference in nonlinear state-space models.

The second category involves relational models. This means that observations may have distinctive internal structure and they may be linked to each other. A novel method called logical hidden Markov model is introduced for analysing sequences of logical atoms, and applied to classifying protein secondary structures. Algorithms for inference, parameter estimation, and structural learning are given. Also, the first graphical model for analysing nonlinear dependencies in relational data, is introduced in the thesis.

Compaction of C-Band Synthetic Aperture Radar Based Sea Ice Information for Navigation in the Baltic Sea

Juha Karvonen

Dissertation for the degree of Doctor of Science in Technology on 8 December 2006.

External examiners:
Markku Hauta-Kasari (University of Joensuu)
Matti Leppäranta (University of Helsinki)
Opponent:
Torbjörn Eltoft (University of Tromsö)

Abstract:
In this work operational sea ice synthetic aperture radar (SAR) data products were improved and developed. The main idea is to deliver the essential SAR-based sea ice information to end-users (typically on ships) in a compact and user-friendly format. The operational systems at Finnish Institute of Marine Research (FIMR) are based on the Canadian SAR-satellite Radarsat-1.

The operational sea ice classification, developed by the author with colleagues, has been further developed. An incidence angle correction algorithm to normalize the backscattering over the SAR incidence angle range for Baltic Sea ice has been developed. The algorithm is based on SAR backscattering statistics over the Baltic Sea.

A SAR segmentation algorithm based on pulse-coupled neural networks has been developed and tested. The parameters have been tuned suitable for the operational data in use at FIMR. The sea ice classification is based on this segmentation and the classification is segment-wise rather than pixel-wise.

To improve distinguishing between sea ice and open water an open water detection algorithm based on segmentation and local autocorrelation has been developed. Also ice type classification based on higher-order statistics and independent component analysis has been studied.

A compression algorithm for compressing sea ice SAR data for visual use has been developed. This algorithm is based on the wavelet decomposition, zero-tree structure and arithmetic coding. Also some properties of the human visual system were utilized.

SAR-based ice thickness estimation has been developed and evaluated. This method uses the ice thickness history derived from digitized ice charts, made daily at the Finnish Ice Service, as its input, and updates this chart based on the novel SAR data. The result is an ice thickness chart representing the ice situation at the SAR acquisition time in higher resolution than in the manually made ice thickness charts. For the evaluation a helicopter-borne ice thickness measuring instrument, based on electromagnetic induction and laser altimeter, was used.

Adaptive combinations of classifiers with application to on-line handwritten character recognition

Matti Aksela

Dissertation for the degree of Doctor of Science in Technology on 29 March, 2007.

External examiners:
David Windridge (University of Surrey)
Jarmo Hurri (University of Helsinki)
Opponent:
Robert P.W. Duin (Delft University of Technology)

Abstract:
Classifier combining is an effective way of improving classification performance. User adaptation is clearly another valid approach for improving performance in a user-dependent system, and even though adaptation is usually performed on the classifier level, also adaptive committees can be very effective. Adaptive committees have the distinct ability of performing adaptation without detailed knowledge of the classifiers. Adaptation can therefore be used even with classification systems that intrinsically are not suited for adaptation, whether that be due to lack of access to the workings of the classifier or simply a classification scheme not suitable for continuous learning.

This thesis proposes methods for adaptive combination of classifiers in the setting of on-line handwritten character recognition. The focal part of the work introduces adaptive classifier combination schemes, of which the two most prominent ones are the Dynamically Expanding Context (DEC) committee and the Class-Confidence Critic Combining (CCCC) committee. Both have been shown to be capable of successful adaptation to the user in the task of on-line handwritten character recognition. Particularly the highly modular CCCC framework has shown impressive performance also in a doubly-adaptive setting of combining adaptive classifiers by using an adaptive committee.

In support of this main topic of the thesis, some discussion on a methodology for deducing correct character labeling from user actions is presented. Proper labeling is paramount for effective adaptation, and deducing the labels from the user's actions is necessary to perform adaptation transparently to the user. In that way, the user does not need to give explicit feedback on the correctness of the recognition results.

Also, an overview is presented of adaptive classification methods for single-classifier adaptation in handwritten character recognition developed at the Laboratory of Computer and Information Science of the Helsinki University of Technology, CIS-HCR. Classifiers based on the CIS-HCR system have been used in the adaptive committee experiments as both member classifiers and to provide a reference level.

Finally, two distinct approaches for improving the performance of committee classifiers further are discussed. Firstly, methods for committee rejection are presented and evaluated. Secondly, measures of classifier diversity for classifier selection, based on the concept of diversity of errors, are presented and evaluated.

The topic of this thesis hence covers three important aspects of pattern recognition: on-line adaptation, combining classifiers, and a practical evaluation setting of handwritten character recognition. A novel approach combining these three core ideas has been developed and is presented in the introductory text and the included publications.

To reiterate, the main contributions of this thesis are: 1) introduction of novel adaptive committee classification methods, 2) introduction of novel methods for measuring classifier diversity, 3) presentation of some methods for implementing committee rejection, 4) discussion and introduction of a method for effective label deduction from on-line user actions, and as a side-product, 5) an overview of the CIS-HCR adaptive on-line handwritten character recognition system.

Dimensionality Reduction for Visual Exploration of Similarity Structures

Jarkko Venna

Dissertation for the degree of Doctor of Science in Technology on 8 June, 2007.

External examiners:
Pasi Fränti (University of Joensuu)
Oleg Okun (University of Oulu)
Opponent:
Michel Verleysen (Université catholique de Louvain)

Abstract:
Visualizations of similarity relationships between data points are commonly used in exploratory data analysis to gain insight on new data sets. Answers are searched for questions like: Does the data consist of separate groups of points? What is the relationship of the previously known interesting data points to other data points? Which points are similar to the points known to be of interest? Visualizations can be used both to amplify the cognition of the analyst and to help in communicating interesting similarity structures found in the data to other people.

One of the main problems faced in information visualization is that while the data is typically very high-dimensional, the display is limited to only two or at most three dimensions. Thus, for visualization, the dimensionality of the data has to be reduced. In general, it is not possible to preserve all pairwise relationships between data points in the dimensionality reduction process. This has lead to the development of a large number of dimensionality reduction methods that focus on preserving different aspects of the data. Most of these methods were not developed to be visualization methods, which makes it hard to assess their suitability for the task of visualizing similarity structures. This problem is made more severe by the lack of suitable quality measures in the information visualization field.

In this thesis a new visualization task, visual neighbor retrieval, is introduced. It formulates information visualization as an information retrieval task. To assess the performance of dimensionality reduction methods in this task two pairs of new quality measures are introduced and the performance of several dimensionality reduction methods are analyzed. Based on the insight gained on the existing methods, three new dimensionality reduction methods (NeRV, fNeRV and LocalMDS) aimed for the visual neighbor retrieval task, are introduced. All three new methods outperform other methods in numerical experiments; they vary in their speed and accuracy.

A new color coding scheme, similarity-based color coding, is introduced in this thesis for visualization of similarity structures, and the applicability of the new methods in the task of creating graph layouts is studied. Finally, new approaches to visually studying the results and convergence of Markov Chain Monte Carlo methods are introduced.

Language Models for Automatic Speech Recognition: Construction and Complexity Control

Vesa Siivola

Dissertation for the degree of Doctor of Science in Technology on 3 September, 2007.

External examiners:
Krister Lindén (University of Helsinki)
Imre Kiss (Nokia Research Center)
Opponent:
Dietrich Klakow (Universität des Saarlandes)

Abstract:
The language model is one of the key components of a large vocabulary continuous speech recognition system. Huge text corpora can be used for training the language models. In this thesis, methods for extracting the essential information from the training data and expressing the information as a compact model are studied.

The thesis is divided in three main parts. In the first part, the issue of choosing the best base modeling unit for the prevalent language modeling method, n-gram language modeling, is examined. The experiments are focused on morpheme-like subword units, although syllables are also tried. Rule-based grammatical methods and unsupervised statistical methods for finding morphemes are compared with the baseline word model. The Finnish cross-entropy and speech recognition experiments show that significantly more efficient models can be created using automatically induced morpheme-like subword units as the basis of the language model.

In the second part, methods for choosing the n-grams that have explicit probability estimates in the n-gram model are studied. Two new methods specialized on selecting the n-grams for Kneser-Ney smoothed n-gram models are presented, one for pruning and one for growing the model. The methods are compared with entropy-based pruning and Kneser pruning. Experiments on Finnish and English text corpora show that the proposed pruning method gives considerable improvements over the previous pruning algorithms for Kneser-Ney smoothed models and also is better than entropy pruned Good-Turing smoothed model. Using the growing algorithm for creating a starting point for the pruning algorithm further improves the results. The improvements in Finnish speech recognition over the other Kneser-Ney smoothed models were significant as well.

To extract more information from the training corpus, words should not be treated as independent tokens. The syntactic and semantic similarities of the words should be taken into account in the language model. The last part of this thesis explores, how these similarities can be modeled by mapping the words into continuous space representations. A language model formulated in the state-space modeling framework is presented. Theoretically, the state-space language model has several desirable properties. The state dimension should determine, how much the model is forced to generalize. The need to learn long-term dependencies should be automatically balanced with the need to remember the short-term dependencies in detail. The experiments show that training a model that fulfills all the theoretical promises is hard: the training algorithm has high computational complexity and it mainly finds local minima. These problems still need further research.

Advances in variable selection and visualization methods for analysis of multivariate data

Timo Similä

Dissertation for the degree of Doctor of Science in Technology on 19 October 2007.

External examiners:
Risto Ritala (Tampere University of Technology)
Patrik Hoyer (University of Helsinki)
Opponent:
Volker Tresp (Siemens Corporate Technology)

Abstract:
This thesis concerns the analysis of multivariate data. The amount of data that is obtained from various sources and stored in digital media is growing at an exponential rate. The data sets tend to be too large in terms of the number of variables and the number of observations to be analyzed by hand. In order to facilitate the task, the data set must be summarized somehow. This work introduces machine learning methods that are capable of finding interesting patterns automatically from the data. The findings can be further used in decision making and prediction. The results of this thesis can be divided into three groups.

The first group of results is related to the problem of selecting a subset of input variables in order to build an accurate predictive model for several response variables simultaneously. Variable selection is a difficult combinatorial problem in essence, but the relaxations examined in this work transform it into a more tractable optimization problem of continuous-valued parameters. The main contribution here is extending several methods that are originally designed for a single response variable to be applicable with multiple response variables as well. Examples of such methods include the well known lasso estimate and the least angle regression algorithm.

The second group of results concerns unsupervised variable selection, where all variables are treated equally without making any difference between responses and inputs. The task is to detect the variables that contain, in some sense, as much information as possible. A related problem that is also examined is combining the two major categories of dimensionality reduction: variable selection and subspace projection. Simple modifications of the multiresponse regression techniques developed in this thesis offer a fresh approach to these unsupervised learning tasks. This is another contribution of the thesis.

The third group of results concerns extensions and applications of the self-organizing map (SOM). The SOM is a prominent tool in the initial exploratory phase of multivariate analysis. It provides a clustering and a visual low-dimensional representation of a set of high-dimensional observations. Firstly, an extension of the SOM algorithm is proposed in this thesis, which is applicable to strongly curvilinear but intrinsically low-dimensional data structures. Secondly, an application of the SOM is proposed to interpret nonlinear quantile regression models. Thirdly, a SOM-based method is introduced for analyzing the dependency of one multivariate data set on another.

Methods for exploring genomic data sets: application to human endogenous retroviruses

Merja Oja

Dissertation for the degree of Doctor of Science in Technology on 14 December 2007.

External examiners:
Juho Rousu (University of Helsinki)
Tero Aittokallio (University of Turku)
Opponent:
Hiroshi Mamitsuka (Kyoto University)

Abstract:
In this thesis exploratory data analysis methods have been developed for analyzing genomic data, in particular human endogenous retrovirus (HERV) sequences and gene expression data. HERVs are remains of ancient retrovirus infections and now reside within the human genome. Little is known about their functions. However, HERVs have been implicated in some diseases. This thesis provides methods for analyzing the properties and expression patterns of HERVs.

Nowadays the genomic data sets are so large that sophisticated data analysis methods are needed in order to uncover interesting structures in the data. The purpose of exploratory methods is to help in generating hypotheses about the properties of the data. For example, by grouping together genes behaving similarly, and hence presumably having similar function, a new function can be suggested for previously uncharacterized genes. The hypotheses generated by exploratory data analysis can be verified later in more detailed studies. In contrast, a detailed analysis of all the genes of an organism would be too time consuming and expensive.

In this thesis self-organizing map (SOM) based exploratory data analysis approaches for visualization and grouping of gene expression profiles and HERV sequences are presented. The SOM-based analysis is complemented with estimates on reliability of the SOM visualization display. New measures are developed for estimating the relative reliability of different parts of the visualization. Furthermore, methods for assessing the reliability of groups of samples manually extracted from a visualization display are introduced.

Finally, a new computational method is developed for a specific problem in HERV biology. Activities of individual HERV sequences are estimated from a database of expressed sequence tags using a hidden Markov mixture model. The model is used to analyze the activity patterns of HERVs.

Theses

Licentiate of Science in Technology

2007

Hiisilä, Heli
Segmentation of time series and sequences using basis representations

Master of Science in Technology

2006

Aaltonen, Leif
Utilization of a specific controller for compensation and control functions in a RFIC of a mobile communication device

Ajanki, Antti
Modeling of gene regulation with context dependent bayesian networks (Geenisäätelyn mallinnus tilanneriippuvilla Bayes-verkoilla)

Etholén, Antti
The automatic adaptation of hip implants into x-ray images. (Lonkan tekonivelen automaattinen sovitus röntgenkuvaan)

Jokela, Matti
Condition monitoring and algorithms for fault detection (Kunnonvalvonta ja vianilmaisualgoritmit)

Korpela, Mikko
Analysis of changes in gene expression time series data

Kyröhonka, Jussi
Evaluation of the use of JavaServer Faces in a portlet-based application environment

Lehtola, Harri
Peer-to-peer based reliability for multicast sessions

Lundqvist, Leo
Deriving a rule set from a large set of data

Mansikkamäki, Mira
Visualization of computer tomography data for dental implant planning (Tietokonetomo-

grafiadatan visualisointi hammasimplanttisuunnittelua varten)

Paukkeri, Mari-Sanna
Potential risks of user's privacy in uticomp environment

Rasinen, Antti
Analysis of Linux evolution using aligned source code segments

Reyhani, Nima
Noise variance estimation for function approximation

Rinnet, Tapio
Extracting stress-related effects from yeast gene expression by canonical correlation analysis

Ruuskanen, Samuli
Outlook for the global delivery management market

Saarela, Atte
Optimization of wood cutting using dynamic programming (Lautojen sahauksen optimointi dynaamisella optimoinnilla)

Sjöberg, Mats
Content-based retrieval of hierarchical objects with PicSOM

Toivonen, Ville-Matti
Monitoring and maintenance framework for a distributed system

Ukkonen, Tomas
Parallel Monte Carlo simulation in spatial analysis (Hajautettu Monte Carlo simulaatio paikkatietoanalyysissä)

Virtanen, Taina
Self-organizing maps in customer segmentation (Asiakaskannan segmentointi menetelmänä itseorganisoivat kartat)

Ye, Yujie
Variant DSP solutions for educational purpose

2007

Aukia, Janne
Bayesian clustering of huge friendship networks

Björkqvist, Mathias
Integrity protection of untrusted storage

Knuuttila, Olli
Gaussian processes in biostatistics: a case study of personal emergency link usage in Hong Kong

Kärnä, Tuomas
Functional data dimensionality reduction for machine learning

Matilainen, Jukka

A probabilistic modeling toolkit for a systems biology software platform

Multanen, Mikko
Outlier detection in cellular network data exploration

Muurinen, Hannes
Video segmentation and shot boundary detection using self-organizing maps

Nikkilä, Raimo
Farm management information system architecture for precision agriculture

Pietiläinen, Anna-Kaisa
Measuring human mobility

Remes, Ulpu
Speaker-based segmentation and adaptation in automatic speech recognition

Schleimer, Jan-Hendrik
Phase synchronisation in superimposed electrophysiological data

Seppä, Jeremias
Control of an interferometrically traceable metrology atomic force microscope for MIKES

Talonen, Jaakko
Fault detection by adaptive process modeling for nuclear power plant

Tergujeff, Renne
Detecting sea-ice ridges from airborne optical colour images using shadow information

Toivola, Janne
Modular specification of dynamic Bayesian networks for time series analysis

Varjokallio, Matti
Subspace methods for Gaussian mixture models in automatic speech recognition

Ylinen, Tuomo
Closed-loop system for adaptative proportional pressure support

I–Adaptive Informatics Research Centre Research Projects

Chapter 1

Introduction

On the first of January 2006, a new Centre of Excellence called the Adaptive Informatics Research Centre (AIRC) started in the Laboratory of Computer and Information Science at Helsinki University of Technology. It followed the tradition of the Neural Networks Research Centre (NNRC), operative from 1994 to 2005 under the national Centre of Excellence status.

Historically, the core function and strength of our research centre has been the ability to analyze and process extensive data sets using our own innovative methods. Up to 2005, our research concentrated on neurocomputing and statistical machine learning algorithms, with a number of applications. In the algorithmic research, we have attained a world class status especially in such unsupervised machine learning methods as the Self-Organizing Map and Independent Component Analysis.

Building on this solid foundation, we have started to apply the knowledge, expertise and tools to advance knowledge in other domains and disciplines. In the AIRC, we take a more goal-oriented, ambitious, and interdisciplinary approach in targeting at the adaptive informatics problem. By adaptive informatics we mean a field of research where automated learning algorithms are used to discover the relevant informative concepts, components, and their mutual relations from large amounts of data. Access to the ever-increasing amounts of available data and its transformation to forms intelligible for the human user is one of the grand challenges in the near future.

The AIRC Centre of Excellence focuses on several adaptive informatics problems. One is the efficient retrieval and processing techniques for text, digital audio and video, and numerical data such as biological and medical measurements, which will create valuable information sources. Another problem area are advanced multimodal natural interfaces. We are building systems that process multimodal contextual information including spoken and written language, images, videos, and explicit and implicit user feedback. Automated semantic processing of such information will facilitate cost-effective knowledge acquisition and knowledge translation without the need to build the descriptions manually. Yet another problem, which we approach together with experts in brain science and molecular biology, is to develop and apply our algorithmic methods to problems in neuroinformatics and bioinformatics.

The Adaptive Informatics methodology that we focus on is to build empirical models of the data by using automated machine learning techniques, in order to make the information usable. The deep expertise on the algorithmic methods, gained over the years, is used to build realistic solutions, starting from the problem requirements. The application domains have been chosen because of our acquired knowledge in some of their core problems, because of their strategic importance in the near future, and because of their mutual interrelations. The algorithms are based on our own core expertise. Future research will

continue to be novel, innovative, as well as inter- and multi-disciplinary, with a specific focus on shared research activities that will have a significant societal impact.

The AIRC Centre of Excellence consists of five interrelated research groups: Algorithms and Methods, Bioinformatics and Neuroinformatics, Multimodal Interfaces, Computational Cognitive Systems, and Adaptive Informatics Applications (see Figure 1).

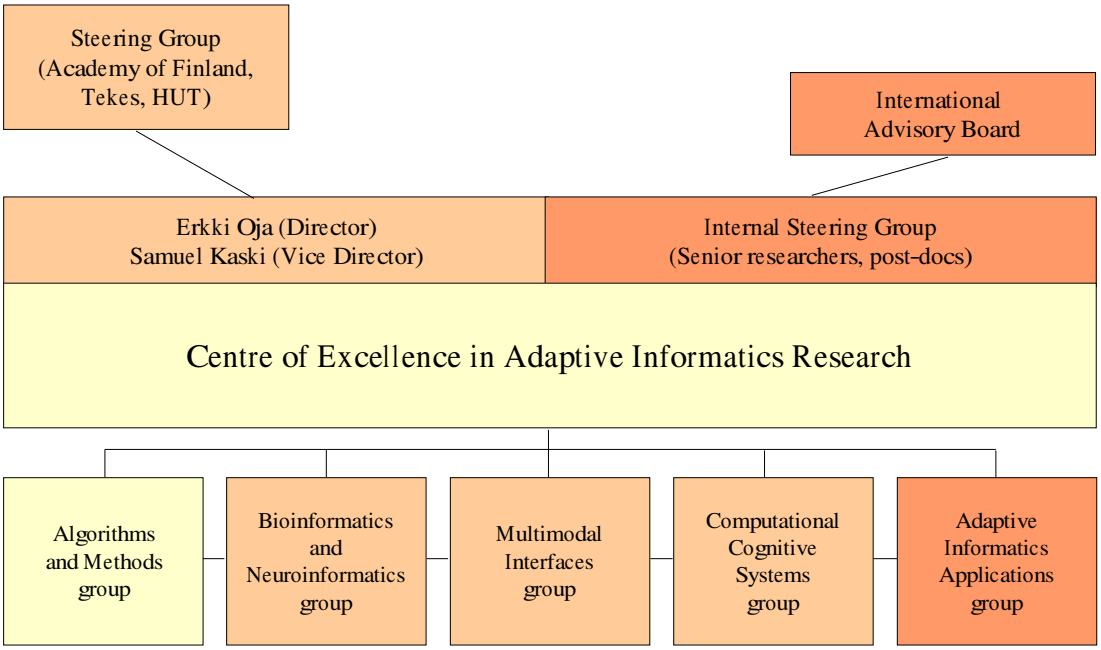


Figure 1.1: *The organization of the AIRC Centre of Excellence*

The Algorithms and Methods group conducts basic algorithmic research in adaptive informatics that relies heavily on computer science, mathematics and statistics, and is partly motivated by the research problems of other groups. In contrast, the groups of Bioinformatics and Neuroinformatics, Multimodal Interfaces and Computational Cognitive Systems form an interdisciplinary research network with shared research interests in life and human sciences. The group of Adaptive Informatics Applications brings the research results into practice together with collaborating enterprises. This inter- and multi-disciplinary diversity facilitates a rich exchange of ideas, knowledge and expertise both within and between research groups. The ideas generated in one research group spark innovative ideas and research methods in other groups. This ability to pool knowledge and resources between groups reduces duplication, saves time, and generates more powerful research methods and results. Altogether, it makes the Centre of Excellence a coherent whole.

Each group has a wide range of national and international collaborators both in Academia and industry. Researcher training, graduate studies, and promotion of creative research is strongly emphasized, following the successful existing traditions.

The present Biennial Report 2006 - 2007 details the individual research projects of the five groups during the first two years of operations of the AIRC. Additional information including demos etc. is available from our Web pages, `www.cis.hut.fi/research`.

Algorithms and Methods

Chapter 2

Bayesian learning of latent variable models

Juha Karhunen, Antti Honkela, Tapani Raiko, Markus Harva, Alexander Ilin, Matti Tornio, Harri Valpola

2.1 Bayesian modeling and variational learning: introduction

Unsupervised learning methods are often based on a generative approach where the goal is to find a latent variable model which explains how the observations were generated. It is assumed that there exist certain latent variables (also called in different contexts source signals, factors, or hidden variables) which have generated the observed data through an unknown mapping. The goal of generative learning is to identify both the latent variables and the unknown generative mapping.

The success of a specific model depends on how well it captures the structure of the phenomena underlying the observations. Various linear models have been popular, because their mathematical treatment is fairly easy. However, in many realistic cases the observations have been generated by a nonlinear process. Unsupervised learning of a nonlinear model is a challenging task, because it is typically computationally much more demanding than for linear models, and flexible models require strong regularization for avoiding overfitting.

In Bayesian data analysis and estimation methods, all the uncertain quantities are modeled in terms of their joint probability distribution. The key principle is to construct the joint posterior distribution for all the unknown quantities in a model, given the data sample. This posterior distribution contains all the relevant information on the parameters to be estimated in parametric models, or the predictions in non-parametric prediction or classification tasks [1, 2].

Denote by $\mathcal{H}$ the particular model under consideration, and by $\boldsymbol{\theta}$ the set of model parameters that we wish to infer from a given data set X. The posterior probability density $p(\boldsymbol{\theta}|X, \mathcal{H})$ of the parameters given the data X and the model $\mathcal{H}$ can be computed from the Bayes' rule

$$p(\boldsymbol{\theta}|X, \mathcal{H}) = \frac{p(X|\boldsymbol{\theta}, \mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})}{p(X|\mathcal{H})} \tag{2.1}$$

Here $p(X|\boldsymbol{\theta}, \mathcal{H})$ is the likelihood of the parameters $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathcal{H})$ is the prior pdf of the parameters, and $p(X|\mathcal{H})$ is a normalizing constant. The term $\mathcal{H}$ denotes all the assumptions made in defining the model, such as the choice of a particular model class and structure, specific noise model, etc.

The parameters $\boldsymbol{\theta}$ of a particular model $\mathcal{H}_i$ are often estimated by seeking the peak value of a probability distribution. The non-Bayesian maximum likelihood (ML) method uses to this end the distribution $p(X|\boldsymbol{\theta}, \mathcal{H})$ of the data, and the Bayesian maximum a posteriori (MAP) method finds the parameter values that maximize the posterior probability density $p(\boldsymbol{\theta}|X, \mathcal{H})$. However, using point estimates provided by the ML or MAP methods is often problematic, because the model order estimation and overfitting (choosing too complicated a model for the given data) are severe problems [1, 2].

Instead of searching for some point estimates, the correct Bayesian procedure is to use all possible models to evaluate predictions and weight them by the respective posterior probabilities of the models. This means that the predictions will be sensitive to regions where the probability mass is large instead of being sensitive to high values of the probability density [3, 2]. This procedure optimally solves the issues related to the model complexity and choice of a specific model $\mathcal{H}_i$ among several candidates. In practice, however, the differences between the probabilities of candidate model structures are often very large, and hence it is sufficient to select the most probable model and use the estimates or predictions given by it.

A problem with fully Bayesian estimation is that the posterior distribution (2.1) has a highly complicated form except for in the simplest problems. Therefore it is too difficult

to handle exactly, and some approximative method must be used. Variational methods form a class of approximations where the exact posterior is approximated with a simpler distribution [4, 2]. In a method commonly known as *Variational Bayes (VB)* [1, 3, 2] the misfit of the approximation is measured by the Kullback-Leibler (KL) divergence between two probability distributions $q(v)$ and $p(v)$. The KL divergence is defined by

$$D(q \parallel p) = \int q(v) \ln \frac{q(v)}{p(v)} dv \tag{2.2}$$

which measures the difference in the probability mass between the densities $q(v)$ and $p(v)$.

A key idea in the VB method is to minimize the misfit between the actual posterior pdf and its parametric approximation using the KL divergence. The approximating density is often taken a diagonal multivariate Gaussian density, because the computations become then tractable. Even this crude approximation is adequate for finding the region where the mass of the actual posterior density is concentrated. The mean values of the Gaussian approximation provide reasonably good point estimates of the unknown parameters, and the respective variances measure the reliability of these estimates.

A main motivation of using VB is that it avoids overfitting which would be a difficult problem if ML or MAP estimates were used. VB method allows one to select a model having appropriate complexity, making often possible to infer the correct number of latent variables or sources. It has provided good estimation results in the very difficult unsupervised (blind) learning problems that we have considered.

Variational Bayes is closely related to information theoretic approaches which minimize the description length of the data, because the description length is defined to be the negative logarithm of the probability. Minimal description length thus means maximal probability. In the probabilistic framework, we try to find the latent variables or sources and the nonlinear mapping which most probably correspond to the observed data. In the information theoretic framework, this corresponds to finding the latent variables or sources and the mapping that can generate the observed data and have the minimum total complexity. This information theoretic view also provides insights to many aspects of learning and helps to explain several common problems [5].

In the following subsections, we first discuss a natural conjugate gradient algorithm which speeds up learning remarkably compared with alternative variational Bayesian learning algorithms. We then briefly present a practical building block framework that can be used to easily construct new models. This work has been for the most part carried out already before the years 2006-2007 covered in this biennial report. After this we consider the difficult nonlinear blind source separation (BSS) problem using our Bayesian methods. This section has been placed into the Bayes chapter instead of the ICA/BSS because the methods used are all Bayesian. This section is followed by variational Bayesian learning of nonlinear state-space models, which are applied to time series prediction, improving inference of states, and stochastic nonlinear model predictive control. After this we consider an approach for non-negative blind source separation, and then principal component analysis in the case of missing values using both Bayesian and non-Bayesian approaches. We then discuss predictive uncertainly and probabilistic relational models. Finally we present applications of the developed Bayesian methods to astronomical data analysis problems. In most of these topics, variational Bayesian learning is used, but for relational models and estimation of time delays in astronomical applications other Bayesian methods are applied.

2.2 Natural conjugate gradient in variational inference

Variational methods for approximate inference in machine learning often adapt a parametric probability distribution to optimize a given objective function. This view is especially useful when applying variational Bayes (VB) to models outside the conjugate-exponential family. For them, variational Bayesian expectation maximization (VB EM) algorithms are not easily available, and gradient-based methods are often used as alternatives.

In previous machine learning algorithms based on natural gradients [6], the aim has been to use maximum likelihood to directly update the model parameters $\boldsymbol{\theta}$ taking into account the geometry imposed by the predictive distribution for data $p(\boldsymbol{X}|\boldsymbol{\theta})$. The resulting geometry is often very complicated as the effects of different parameters cannot be separated and the Fisher information matrix is relatively dense.

Recently, in [7], we propose using natural gradients for free energy minimisation in variational Bayesian learning using the simpler geometry of the approximating distributions $q(\boldsymbol{\theta}|\boldsymbol{\xi})$. Because the approximations are often chosen to minimize dependencies between different parameters $\boldsymbol{\theta}$, the resulting Fisher information matrix with respect to the variational parameters $\boldsymbol{\xi}$ will be mostly diagonal and hence easy to invert.

While taking into account the structure of the approximation, plain natural gradient in this case ignores the structure of the model and the global geometry of the parameters $\boldsymbol{\theta}$. This can be addressed by using conjugate gradients. Combining the natural gradient search direction with a conjugate gradient method yields our proposed *natural conjugate gradient (NCG)* method, which can also be seen as an approximation to the fully Riemannian conjugate gradient method.

The NCG algorithm was compared against conjugate gradient (CG) and natural gradient (NG) algorithms in learning a nonlinear state-space model [8]. The results for a number of datasets ranging from 200 to 500 samples of 21 dimensional speech spectrograms can be seen in Figure 2.1. The plain CG and NG methods were clearly slower than others and the maximum runtime of 24 hours was reached by most CG and some NG runs. NCG was clearly the fastest algorithm with the older heuristic method of [8] between these extremes. The results with a larger data set are very similar with NCG outperforming all alternatives by a factor of more than 10.

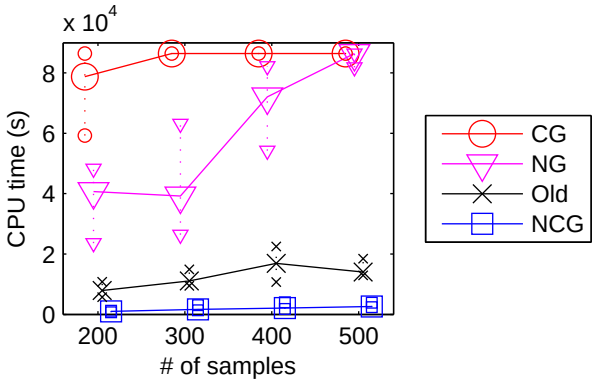


Figure 2.1: Convergence speed of the natural conjugate gradient (NCG), the natural gradient (NG) and the conjugate gradient (CG) methods as well as the heuristic algorithm (Old) with different data sizes. The lines show median times with 25 % and 75 % quantiles shown by the smaller marks. The times were limited to at most 24 hours, which was reached by a number of simulations.

The experiments in [7] show that the natural conjugate gradient method outperforms both conjugate gradient and natural gradient methods by a large margin. Considering univariate Gaussian distributions, the regular gradient is too strong for model variables with small posterior variance and too weak for variables with large posterior variance. The posterior variance of latent variables is often much larger than the posterior variance of model parameters and the natural gradient takes this into account in a very natural manner.

2.3 Building blocks for variational Bayesian learning

In graphical models, there are lots of possibilities to build the model structure that defines the dependencies between the parameters and the data. To be able to manage the variety, we have designed a modular software package using C++/Python called the Bayes Blocks [9]. The theoretical framework on which it is based on was published in [10] and a description of the software package was published in [11].

The design principles for Bayes Blocks have been the following. Firstly, we use standardized building blocks that can be connected rather freely and can be learned with local learning rules, i.e. each block only needs to communicate with its neighbors. Secondly, the system should work with very large scale models. We have made the computational complexity linear with respect to the number of data samples and connections in the model.

The building blocks include Gaussian variables, summation, multiplication, nonlinearity, mixture-of-Gaussians, and rectified Gaussians. Each of the blocks can be a scalar or a vector. Variational Bayesian learning provides a cost function which can be used for updating the variables as well as optimizing the model structure. The derivation of the cost function and learning rules is automatic which means that the user only needs to define the connections between the blocks. Examples of structures which can be build using the Bayes Blocks library can be found in Figure 2.2.

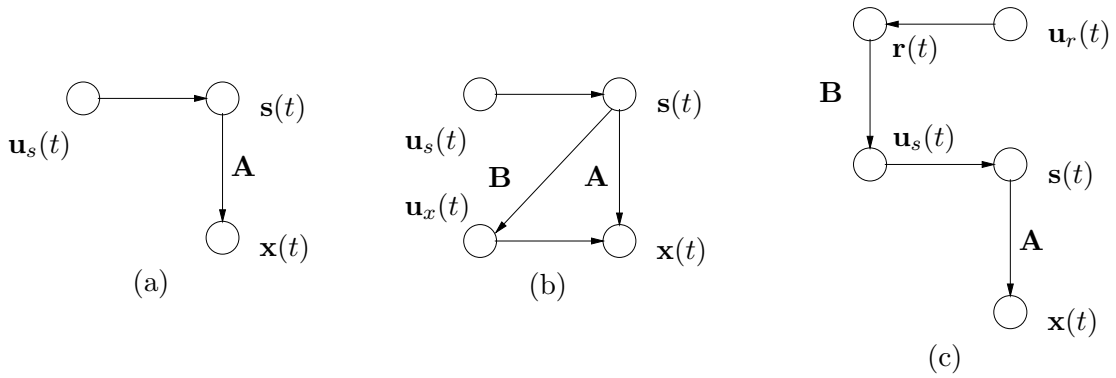


Figure 2.2: Various model structures utilizing variance nodes. Observations are denoted by $\mathbf{x}$, linear mappings by $\mathbf{A}$ and $\mathbf{B}$, sources by $\mathbf{s}$ and $\mathbf{r}$, and variance nodes by $\mathbf{u}$.

2.4 Nonlinear BSS and ICA

A fundamental difficulty in the nonlinear blind source separation (BSS) problem and even more so in the nonlinear independent component analysis (ICA) problem is that they provide non-unique solutions without extra constraints, which are often implemented by using a suitable regularization. Our approach to nonlinear BSS uses Bayesian inference methods for estimating the best statistical parameters, under almost unconstrained models in which priors can be easily added.

We have applied variational Bayesian learning to nonlinear factor analysis (FA) and BSS where the generative mapping from sources to data is not restricted to be linear. The general form of the model is

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_f) + \mathbf{n}(t) \tag{2.3}$$

This can be viewed as a model about how the observations were generated from the sources. The vectors $\mathbf{x}(t)$ are observations at time t, $\mathbf{s}(t)$ are the sources, and $\mathbf{n}(t)$ the noise. The function $\mathbf{f}(\cdot)$ is a mapping from source space to observation space parametrized by $\boldsymbol{\theta}_f$.

In an earlier work [13] we have used multi-layer perceptron (MLP) network with tanh-nonlinearities to model the mapping $\mathbf{f}$:

$$\mathbf{f}(\mathbf{s}; \mathbf{A}, \mathbf{B}, \mathbf{a}, \mathbf{b}) = \mathbf{B} \tanh(\mathbf{A}\mathbf{s} + \mathbf{a}) + \mathbf{b}\,. \tag{2.4}$$

The mapping $\mathbf{f}$ is thus parameterized by the matrices $\mathbf{A}$ and $\mathbf{B}$ and bias vectors $\mathbf{a}$ and $\mathbf{b}$. MLP networks are well suited for nonlinear FA and BSS. First, they are universal function approximators which means that any type of nonlinearity can be modeled by them in principle. Second, it is easy to model smooth, nearly linear mappings with them. This makes it possible to learn high dimensional nonlinear representations in practice.

An important special case of general nonlinear mixtures in (2.3) is a post-nonlinear (PNL) mixing model. There linear mixing is followed by component-wise nonlinearities acting on each output independently of the others:

$$x_i(t) = f_i\left[\mathbf{a}_i^T \mathbf{s}(t)\right] + n_i(t) \qquad i = 1, \ldots, n \tag{2.5}$$

Such models are plausible in applications where linearly mixed signals are measured by sensors with nonlinear distortions f_i. The nonlinearities f_i can also be modelled by MLP networks.

Identification of models (2.3) or (2.5) assuming Gaussianity of sources $\mathbf{s}(t)$ helps to find a compact representation of the observed data $\mathbf{x}(t)$. Nonlinear BSS can be achieved by performing a linear rotation of the found sources using, for example, a linear ICA technique.

The paper [12] presents our recent developments on nonlinear FA and BSS. A more accurate linearization increases stability of the algorithm in cases with a large number of sources when the posterior variances of the last weak sources are typically large. A hierarchical nonlinear factor analysis (HNFA) model using the building blocks presented in Section 2.3 is applicable to larger problems than the MLP based method, as the computational complexity is linear with respect to the number of sources. Estimating the PNL factor analysis model in (2.5) using variational Bayesian learning helps achieve separation of signals in very challenging BSS problems.

2.5 Nonlinear state-space models

In many cases, measurements originate from a dynamical system and form a time series. In such instances, it is often useful to model the dynamics in addition to the instantaneous observations. We have used rather general nonlinear models for both the data (observations) and dynamics of the sources (latent variables) [8]. This results in a state-space model where the sources can be interpreted as the internal state of the underlying generative process.

The general form of our nonlinear model for the generative mapping from the source (latent variable) vector $\mathbf{s}(t)$ to the data (observation) vector $\mathbf{x}(t)$ at time t is the same as in Eq. (2.3):

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_f) + \mathbf{n}(t)\,. \tag{2.6}$$

The dynamics of the sources can be modelled by another nonlinear mapping, which leads to a source model [8]

$$\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t-1), \boldsymbol{\theta}_g) + \mathbf{m}(t)\,, \tag{2.7}$$

where $\mathbf{s}(t)$ are the sources (states) at time t, $\mathbf{m}$ is the Gaussian noise, and $\mathbf{g}(\cdot)$ is a vector containing as its elements the nonlinear functions modelling the dynamics.

As for the static models presented in Sec. 2.4, the nonlinear functions are modelled by MLP networks. The mapping $\mathbf{f}$ has the same functional form (2.4). Since the states in dynamical systems are often slowly changing, the MLP network for mapping $\mathbf{g}$ models the change in the value of the source:

$$\mathbf{g}(\mathbf{s}(\mathbf{t}-\mathbf{1})) = \mathbf{s}(t-1) + \mathbf{D}\tanh[\mathbf{C}\mathbf{s}(t-1) + \mathbf{c}] + \mathbf{d}\,. \tag{2.8}$$

The dynamic mapping $\mathbf{g}$ is thus parameterized by the matrices $\mathbf{C}$ and $\mathbf{D}$ and bias vectors $\mathbf{c}$ and $\mathbf{d}$.

Estimation of the arising state-space model is rather involved, and it is discussed in detail in our earlier paper [8]. An important advantage of the proposed nonlinear state-space method (NSSM) is its ability to learn a high-dimensional latent source space. We have also reasonably solved computational and over-fitting problems which have been major obstacles in developing this kind of unsupervised methods thus far. Potential applications for our method include prediction and process monitoring, control and identification. MATLAB software packages are available for both the static model (2.3)-(2.4) (under the name nonlinear factor analysis) and the dynamic model (2.7)-(2.8) (under the name nonlinear dynamical factor analysis) on the home page of our Bayes group [14].

Time series prediction

Traditionally, time series prediction is done using models based directly on the past observations of the time series. Perhaps the two most important classes of neural network based solutions used for nonlinear prediction are feedforward autoregressive neural networks and recurrent autoregressive moving average neural networks [15]. However, instead of modelling the system based on past observations, it is also possible to model the same information in a more compact form with a state-space model.

We have used the nonlinear state-space model and method [8] described in the beginning of this section to model a time series. The primary goal in the paper [16] was to apply our NSSM method and software [14] to the time series prediction task as a black box tool. The details of this application are given in [16].

We applied the NSSM method to the prediction of the nonlinear scalar time series provided by the organizers of the ESTSP'07 symposium. The original time series containing 875 samples is shown in Figure 2.3. It seems to be strongly periodic with a period of

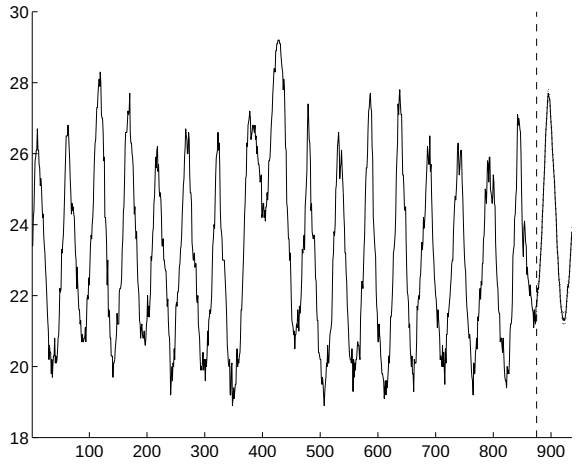


Figure 2.3: The original time series and the predicted 61 next time steps.

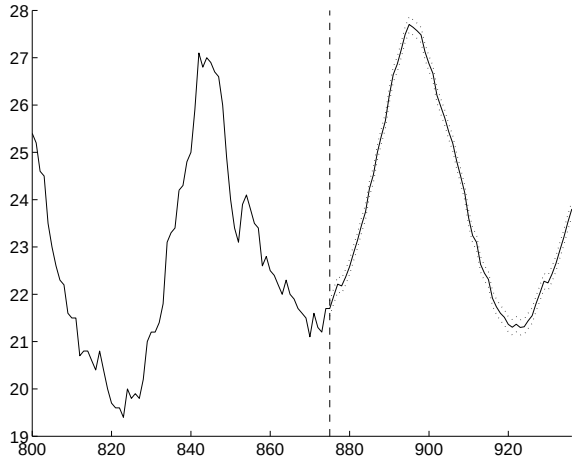


Figure 2.4: Bottom: The original time series starting from time instant 800 and the predicted 61 next time steps.

approximately 52 samples. Figure 2.3 shows also the predicted 61 next time steps, and Figure 2.4 in more detail the original time series starting from time instant 800 and the predicted 61 next time steps. The dotted lines in both figures represent pseudo 95 % confidence intervals. These intervals are, however, smaller than in reality as the variance caused by the innovation is ignored [16].

Improving state inference

The problem of state inference involves finding the source vectors $\mathbf{s}(t-1)$ given the data and the model. While this is an easier problem than finding both the model and the sources, it is more time critical, since it must often be computed in real-time. While the algorithm in [8] can be used for inference, it is very slow because of the slow flow of information through the time series. Standard algorithms based on extensions of the Kalman smoother work rather well in general, but may fail to converge when estimating the states over a long gap or when used together with learning the model.

When updates are done locally, information spreads around slowly because the states of different time slices affect each other only between updates. It is possible to predict this interaction by a suitable approximation. In [17], we derived a novel update algorithm

for the posterior mean of the states by replacing partial derivatives of the cost function with respect to state means $\overline{\mathbf{s}}(t)$ by (approximated) total derivatives

$$\frac{d\mathcal{C}_{\mathrm{KL}}}{d\overline{\mathbf{s}}(t)} = \sum_{\tau=1}^{T} \frac{\partial \mathcal{C}_{\mathrm{KL}}}{\partial \overline{\mathbf{s}}(\tau)} \frac{\partial \overline{\mathbf{s}}(\tau)}{\partial \overline{\mathbf{s}}(t)}. \tag{2.9}$$

They can be computed efficiently using the chain rule and dynamic programming, given that we can approximate the terms $\partial\overline{\mathbf{s}}(t)/\partial\overline{\mathbf{s}}(t-1)$ and $\partial\overline{\mathbf{s}}(t)/\partial\overline{\mathbf{s}}(t+1)$.

This is how we approximated the required partial derivatives. The posterior distribution of the state $\mathbf{s}(t)$ can be factored into three potentials, one from $\mathbf{s}(t-1)$ (the past), one from $\mathbf{s}(t+1)$ (the future), and one from $\mathbf{x}(t)$ (the observation). We linearized the nonlinear mappings so that the three potentials become Gaussian. Then also the posterior of $\mathbf{s}(t)$ becomes Gaussian with a mean that is the weighted average of the means of the three potentials, where the weights are the inverse (co)variances of the potentials. A change in the mean of a potential results in a change of the mean of the posterior inversely proportional to their (co)variances.

Experimental comparison in [17] showed that the proposed algorithm worked reliably and fast. The algorithms from the Kalman family (IEKS and IUKS) were fast, too, but they also suffered from stability problems when gaps of 30 consequtive missing observations were introduced into the data. Basic particle smoother performed very poorly compared to the iterative algorithms. It should be noted that many different schemes exists to improve the performance of particle filters.

Stochastic nonlinear model-predictive control

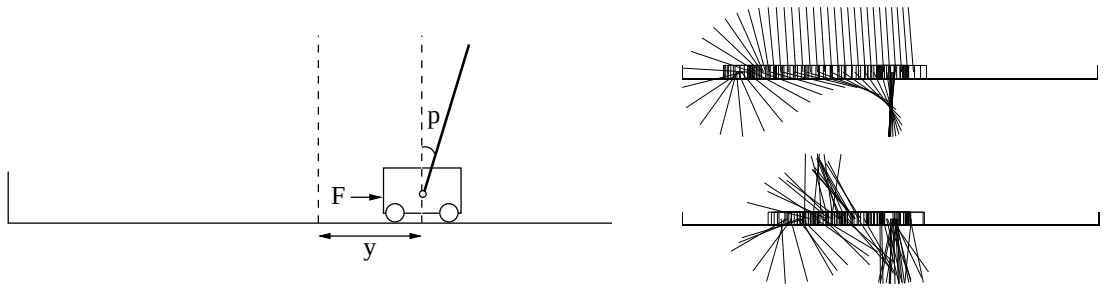


Figure 2.5: Left: The cart-pole system. The goal is to swing the pole to an upward position and stabilize it without hitting the walls. The cart can be controlled by applying a force to it. Top left: The pole is successfully swinged up by moving first to the left and then right. Bottom right: Our controller works quite reliably even in the presence of serious observation noise.

In [18], we studied such a system combining variational Bayesian learning of an unknown dynamical system with nonlinear model-predictive control. For being able to control the dynamical system, control inputs are added to the nonlinear state-space model. Then we can use stochastic nonlinear model-predictive control, which is based on optimising control signals based on maximising a utility function.

Figure 2.5 shows simulations with a cart-pole swing-up task. The results confirm that selecting actions based on a state-space model instead of the observation directly has many benefits: First, it is more resistant to noise because it implicitly involves filtering. Second, the observations (without history) do not always carry enough information about the system state. Third, when nonlinear dynamics are modelled by a function approximator such

as an multilayer perceptron network, a state-space model can find such a representation of the state that it is more suitable for the approximation and thus more predictable.

Continuous-time modeling

In [19], we have outlined an extension of the discrete-time variational Bayesian NSSM of [8] to continuous-time systems and presented preliminary experimental results with the method. Evaluation of the method with larger and more realistic examples is a very important item of further work. The main differences between continuous-time and discrete-time variational NSSMs are the different method needed to evaluate the predictions of the states and the different form of the dynamical noise or innovation.

2.6 Non-negative blind source separation

In linear factor analysis (FA) [20], the observations are modeled as noisy linear combinations of a set of underlying sources or factors. When the level of noise is low, FA reduces to principal component analysis (PCA). Both FA and PCA are insensitive to orthogonal rotations, and, as such, cannot be used for blind source separation except in special cases. There are several ways to solve the rotation indeterminacy. One approach is to assume the sources independent, which in low noise leads to independent component analysis. Another approach, the one discussed in this section, is to constrain the sources to be non-negative.

Non-negativity constraints in linear factor models have received a great deal of interest in a number of problem domains. In the variational Bayesian framework, positivity of the factors can be achieved by putting a non-negatively supported prior on them. The rectified Gaussian distribution is particularly convenient, as it is conjugate to the Gaussian likelihood arising in the FA model. Unfortunately, this solution has a technical limitation: the location parameter of the prior has to be fixed to zero; otherwise the potentials of both the location and the scale parameter become awkward.

To evade the above mentioned problems, the model is reformulated using rectification nonlinearities. This can be expressed in the form of Eq. (2.4) using the following nonlinearity

$$\mathbf{f}(\mathbf{s};\, \mathbf{A}) = \mathbf{A}\, \mathbf{cut}(\mathbf{s}) \tag{2.10}$$

where **cut** is the componentwise rectification (or cut) function such that $[\mathbf{cut}(\mathbf{s})]_i = \max(s_i, 0)$. In [21], a variational learning procedure was derived for the proposed model and it was shown that it indeed overcomes the problems that exist with the related approaches (see Figure 2.6 for a controlled experiment). In Section 2.10 an application of the method to the analysis of galaxy spectra is presented. There the underlying sources were such that the zero-location rectified Gaussian prior was highly inappropriate, which motivated the development of the proposed approach.

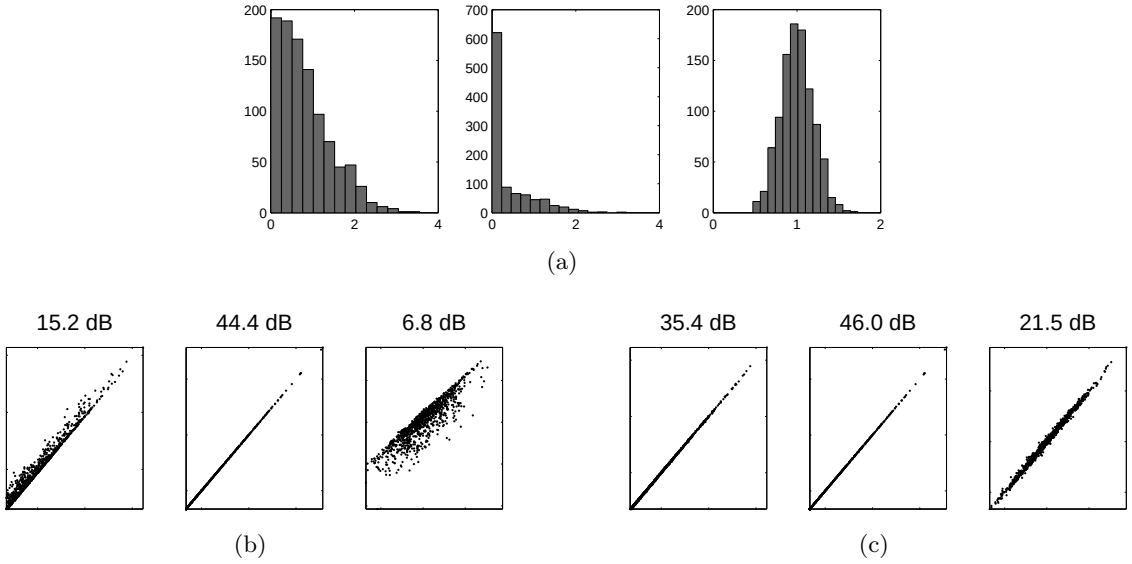


Figure 2.6: (a) The histograms of the true sources to be recovered. (b) and (c) The estimated sources plotted against the true sources with the signal-to-noise ratios printed above each plot. In (b), rectified Gaussian priors have been used for the sources. In (c), the proposed approach employing rectification nonlinearities has been used.

2.7 PCA in the presence of missing values

Principal component analysis (PCA) is a classical data analysis technique. Some algorithms for PCA scale better than others to problems with high dimensionality. They also differ in the ability to handle missing values in the data. In our recent papers [22, 23], a case is studied where the data are high-dimensional and a majority of the values are missing.

In the case of very sparse data, overfitting becomes a severe problem even in simple linear models such as PCA. Regularization can be provided using the Bayesian approach by introducing prior for the model parameters. The PCA model can then be identified using, for example, maximum a posteriori estimates (regularized PCA) or variational Bayesian (VB) learning. We study both approaches in the papers [22, 23].

The proposed learning algorithm is based on speeding up a simple principal subspace rule in which the model parameters are updated as

$$\theta_i \leftarrow \theta_i - \gamma \left(\frac{\partial^2 C}{\partial \theta_i^2} \right)^{-\alpha} \frac{\partial C}{\partial \theta_i}, \tag{2.11}$$

where α is a control parameter that allows the learning algorithm to vary from the standard gradient descent ($\alpha = 0$) to the diagonal Newton's method ($\alpha = 1$). These learning rules can be used for standard PCA learning and extended to regularized PCA and variational Bayesian (VB) PCA.

The algorithms were tested on the Netflix problem (`http://www.netflixprize.com/`), which is a task of predicting preferences (or producing personal recommendations) by using other people's preferences. The Netflix problem consists of movie ratings given by 480189 customers to 17770 movies. There are 100480507 ratings from 1 to 5 given, and the task is to predict 2817131 other ratings among the same group of customers and movies. 1408395 of the ratings are reserved for validation. Thus, 98.8% of the values are missing. We tried to find 15 principal components from the data using a number of methods. The results confirm that the proposed speed-up procedure is much faster than any of the compared methods, and that VB-PCA method provides more accurate predictions for new data than traditional PCA or simple regularized PCA (see Fig. 2.7).

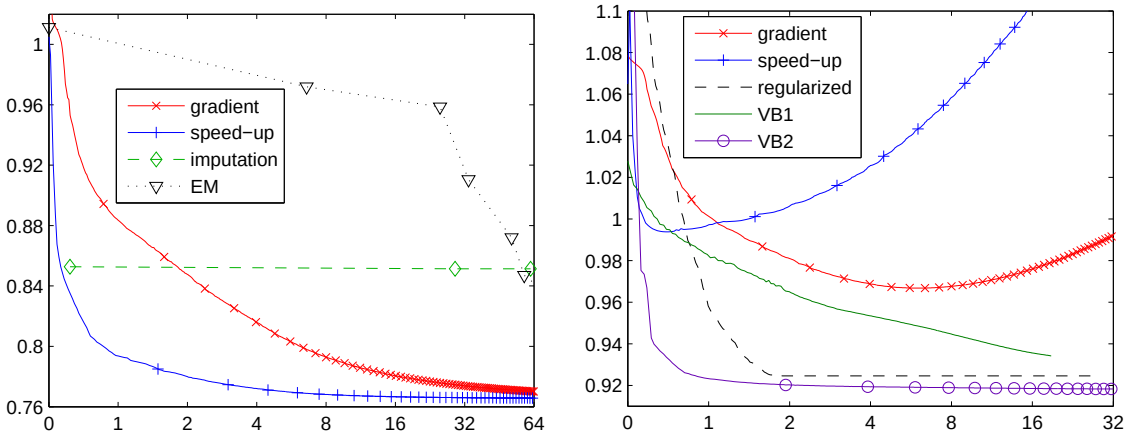


Figure 2.7: *Left:* Training error against computation time in hours in the Netflix problem for unregularized PCA algorithm based on gradient descent and the proposed speed-up. Two alternative methods are shown for comparison. *Right:* The error on test data for the two versions of unregularized PCA, regularized PCA and two variants of variational Bayesian PCA. The time scale is linear below 1 and logarithmic above 1.

2.8 Predictive uncertainty

In standard regression, we seek to predict the value of a response variable based on a set of explanatory variables. Here, the term *predictive uncertainty* is used to refer to a task similar to regression with the exception that we predict not only the mean outcome of the response variable, but also the uncertainty related to its value. For example, consider predicting the concentration of an air pollutant in a city, based on meteorological conditions measured some time in advance. In this task it is the extreme events, namely those occasions when the concentration of the air pollutant rises over a certain threshold, that are interesting. If the conditional distribution of the response variable is not tightly concentrated around its mean value, the mean value by itself will be a poor indicator of the extreme events occurring, and hence predictions based on those alone might lead to policies with ill consequences.

In [26], a method for predictive uncertainty is presented. The method is based on conditioning the scale parameter of the noise process on the explanatory variables and then using MLP networks to model both the location and the scale of the output distribution. The model can be summarised as

$$\begin{aligned} y_t &\sim N(f(\mathbf{x}_t, \boldsymbol{\theta}_y), e^{-u_t}) \\ u_t &\sim N(f(\mathbf{x}_t, \boldsymbol{\theta}_u), \tau^{-1}) \end{aligned} \tag{2.12}$$

Above, y_t is the response variable and $\mathbf{x}_t$ is the vector of explanatory variables. The function f, representing the MLP network, has essentially the same form as in Eq. (2.4). When the latent variable u_t is marginalised out of the model the predictive distribution for y_t becomes super-Gaussian. The extent to which this happens depends on the uncertainty in u_t as measured by the precision parameter τ which is adapted in the learning process. This adaptive nongaussianity of the predictive distribution is highly desirable as then the uncertainty in the scale parameter can be accommodated by making the predictive distribution more robust.

The problem with heteroscedastic models is that learning them using simple methods can be difficult as overfitting becomes a serious concern. Variational Bayesian (VB) methods can, however, largely avoid these problems. Unfortunately, VB methods for nonlinear models, such as that in Eq. (2.12), become involved both in analytic as well as in computational terms. Therefore the learning algorithm in [26] is based on the slightly weaker approximation technique, the variational EM algorithm, and only the "important" parameters have distributional estimates. These parameters include the latent variables u_t, the precision parameter, and the second layer weights of the MLPs. The rest of the parameters, that is, the first layer weights of the MLPs, have point estimates only.

The method summarized in this section was applied to all four datasets in the 'Predictive uncertainty in environmental modelling' competition held at World Congress on Computational Intelligence 2006. The datasets varied in dimensionality from one input variable to 120 variables. The proposed method performed well with all the datasets where heteroscedasticity was an important component being the overall winner of the competition.

2.9 Relational models

In the previous sections, we have studied models belonging to two categories: static and dynamic. In static modeling, each observation or data sample is independent of the others. In dynamic models, the dependencies between consecutive observations are modeled. A generalization of both types of models is that the relations are described in the data itself, that is, each observation might have a different structure.

Logical hidden Markov models

Many real-world sequences such as protein secondary structures or shell logs exhibit rich internal structures. In [24], we have proposed logical hidden Markov models as one solution. They deal with logical sequences, that is, sequences over an alphabet of logical atoms. This comes at the expense of a more complex model selection problem. Indeed, different abstraction levels have to be explored. Logical hidden Markov models (LOHMMs) upgrade traditional hidden Markov models to deal with sequences of structured symbols in the form of logical atoms, rather than characters. Our recent paper [24] formally introduces LOHMMs and presents solutions to the three central inference problems for LOHMMs: evaluation, most likely hidden state sequence, and parameter estimation. The resulting representation and algorithms are experimentally evaluated on problems from the domain of bioinformatics (see Figure 2.8).

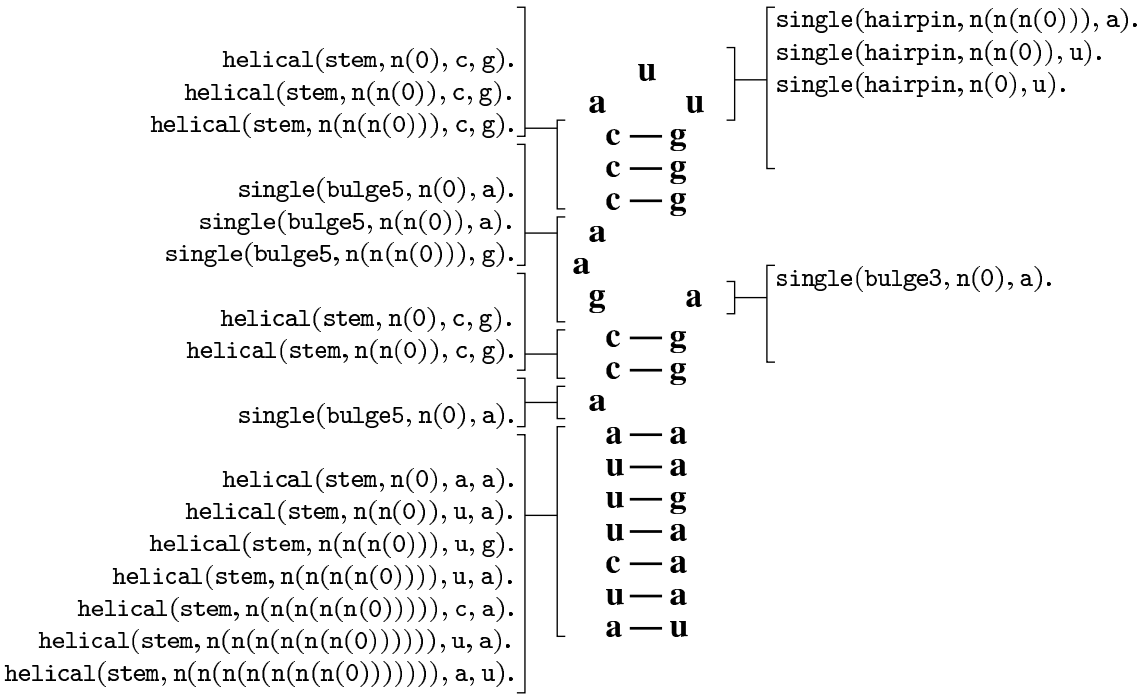


Figure 2.8: Representation of mRNA signal structures as a sequence of logical atoms to be analyzed with a logical hidden Markov model.

Higher order statistics in play-out analysis

A second relational study involves game playing. There is a class of board games called connection games for which traditional artificial intelligence approach does not provide a good computer player. For such games, it is an interesting option to play out the game

from the current state to the end many times randomly. Play-outs provide statistics that can be used for selecting the best move. In [25], we introduce a method that selects relevant patterns of moves to collect higher order statistics. Play-out analysis avoids the horizon effect of regular game-tree search. The proposed method is especially effective when the game can be decomposed into a number of subgames. Preliminary experiments on the board games of Hex and Y are reported in [25].

2.10 Applications to astronomy

Two astronomical applications are discussed in this section: analysis of galaxy spectra and estimation of time delays in gravitational lensing.

Analysis of galaxy spectra

We have applied rectified factor analysis [21] described in Section 2.6 to the analysis of real stellar population spectra of elliptical galaxies. Ellipticals are the oldest galactic systems in the local universe and are well studied in physics. The hypothesis that some of these old galactic systems may actually contain young components is relatively new. Hence, we have investigated whether a set of stellar population spectra can be decomposed and explained in terms of a small set of unobserved spectral prototypes in a data driven but yet physically meaningful manner. The positivity constraint is important in this modelling application, as negative values of flux would not be physically interpretable.

Using a set of 21 real stellar population spectra, we found that they can indeed be decomposed to prototypical spectra, especially to a young and old component [27]. Figure 2.9 shows one spectrum and its decomposition to these two components. The right subfigure shows the ages of the galaxies, known from a detailed astrophysical analysis, plotted against the first weight of the mixing matrix. The plot clearly shows that the first component corresponds to a galaxy containing a significant young stellar population.

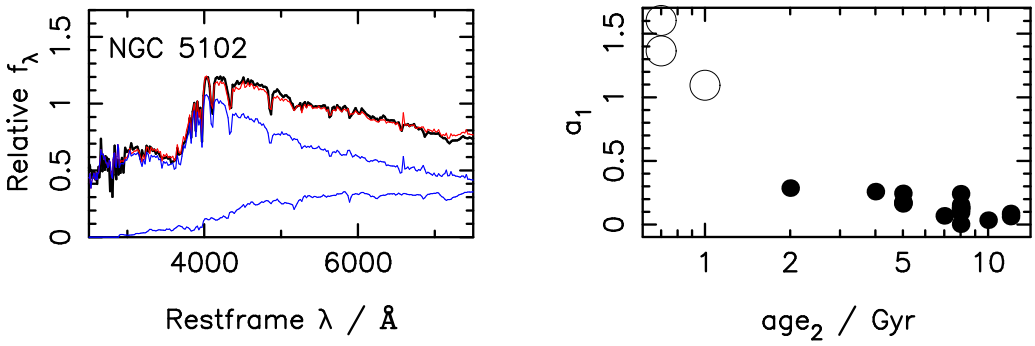


Figure 2.9: Left: the spectrum of a galaxy with its decomposition to a young and old component. Right: the age of the dominating stellar population against the mixing coefficient of the young component.

Estimation of time delays in gravitational lensing

Gravitational lensing occurs when the light coming from a distant bright source is bent by the gravitational potential of an intermediate galaxy such that several images of the source are observed (see the left panel of Figure 2.10 for an example system). Relativistic effects and the different lengths of the paths affect the time it takes for the photons originating from the source to travel to the observer. This is perceived as a delay in the intensity variations between the images (see the right panel of Figure 2.10). The significance of estimating the delays in such systems stems from the early observation that they can be used in determining important cosmological quantities [28].

The delay estimation problem is difficult for various reasons. The main challenge is the uneven sampling rate, as the sampling times are determined by factors one cannot control such as observing conditions and scheduling. The signal-to-noise ratio in the

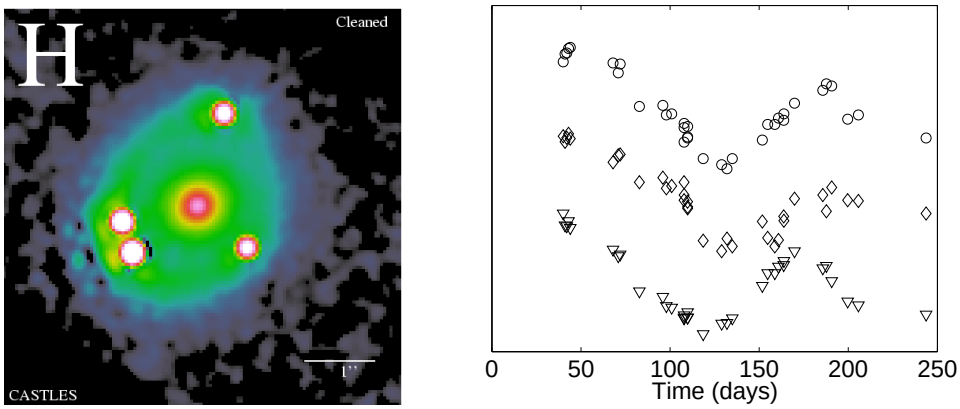


Figure 2.10: Left: The four images of PG1115+080. Right: The corresponding intensity measurements (the two images closest to each other are merged).

observations is often poor too, although this varies somewhat between datasets. Classical delay estimation methods usually rely on the cross-correlation function which is easy to evaluate between regularly sampled signals. The obvious way to attack the problem with unevenly sampled signals would then be to interpolate them appropriately to obtain evenly sampled signals and then apply the cross correlation method. But with all the gaps and the noise in the data, the interpolation can introduce spurious features to the data which make the cross-correlation analysis go awry [29].

In [30, 31], a method for estimating the delay between irregularly sampled signals is presented. Since interpolation on the gappy and noisy data can be venturesome, that is avoided. Instead the two observed signals, $x_1(t)$ and $x_2(t)$, are postulated to have been emitted from the same latent source signal $s(t)$, the observation times being determined by the actual sampling times and the delay. The source is then assumed to follow the Wiener process: $s(t_{i+1}) - s(t_i) \sim N(0, [(t_{i+1} - t_i)\,\sigma]^2)$. This prior encodes the notion of "slow variability" into the model which is an assumption implicitly present in many of the other methods as well. The model is estimated using exact marginalization, which leads to a specific type of Kalman-filter, combined with the Metropolis-Hastings algorithm.

We have used the proposed method to determine the delays in several gravitational lensing systems. Controlled comparisons againts other methods cannot, however, be done with real data as the true delays are unknown to us. Instead, artificial data, where the ground truth is known, must be used. Figure 2.11 shows the performance of several methods in an artificial setting.

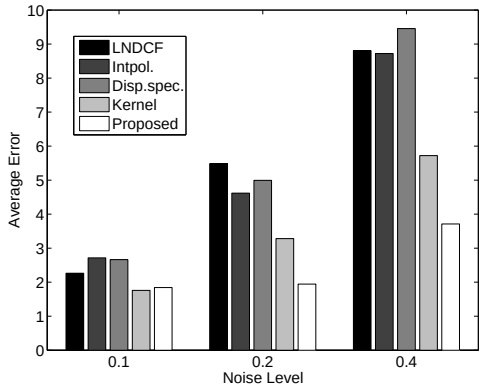


Figure 2.11: Average errors of the methods for three groups of datasets.

References

[1] D. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[2] C. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[3] H. Lappalainen and J. Miskin. Ensemble learning. In M. Girolami, editor, *Advances in Independent Component Analysis*, Springer, 2000, pages 75–92.

[4] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In M. Jordan, editor, *Learning in Graphical Models*, MIT Press, 1999, pages 105–161.

[5] A. Honkela and H. Valpola. Variational learning and bits-back coding: an information-theoretic view to Bayesian learning. *IEEE Transactions on Neural Networks*, 15(4):267–282, 2004.

[6] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[7] A. Honkela, M. Tornio, T. Raiko, and J. Karhunen. Natural conjugate gradient in variational inference. In *Proc. of the 14th Int. Conf. on Neural Information Processing (ICONIP 2007)*, Kitakyushu, Japan, November 2007.

[8] H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692, 2002.

[9] H. Valpola, A. Honkela, M. Harva, A. Ilin, T. Raiko, and T. Östman. Bayes Blocks software library. `http://www.cis.hut.fi/projects/bayes/software/`, 2003.

[10] T. Raiko, H. Valpola, M. Harva, and J. Karhunen. Building blocks for variational Bayesian learning of latent variable models. *Journal of Machine Learning Research*, Vol. 8, pp. 155-201, January 2007.

[11] A. Honkela, M. Harva, T. Raiko, H. Valpola, and J. Karhunen. Bayes Blocks: A Python toolbox for variational Bayesian learning. *NIPS2006 Workshop on Machine Learning Open Source Software*, Whistler, B.C., Canada, 2006.

[12] A. Honkela, H. Valpola, A. Ilin and J. Karhunen. Blind separation of nonlinear mixtures by variational Bayesian learning. *Digital Signal Processing*, Vol. 17, No 2, pp. 914–934, 2007.

[13] H. Lappalainen and A. Honkela. Bayesian nonlinear independent component analysis by multi-layer perceptrons. In Mark Girolami, editor, *Advances in Independent Component Analysis*, pages 93–121. Springer-Verlag, Berlin, 2000.

[14] Home page of our Bayes group: `http://www.cis.hut.fi/projects/bayes/`.

[15] A. Trapletti, *On Neural Networks as Statistical Time Series Models.* PhD Thesis, Technische Universität Wien, 2000.

[16] M. Tornio, A. Honkela, and J. Karhunen. Time series prediction with variational Bayesian nonlinear state-space models. In *Proc. European Symp. on Time Series Prediction (ESTSP'07)*, pages 11–19, Espoo, Finland, February 2007.

[17] T. Raiko, M. Tornio, A. Honkela, and J. Karhunen. State inference in variational Bayesian nonlinear state-space models. In *Proc. of the 6th Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA 2006)*, pages 222–229, Charleston, South Carolina, USA, March 2006.

[18] M. Tornio and T. Raiko. Variational Bayesian approach for nonlinear identification and control. In *Proc. of the IFAC Workshop on Nonlinear Model Predictive Control for Fast Systems, NMPC FS06*, pp. 41-46, Grenoble, France, October 9-11, 2006.

[19] A. Honkela, M. Tornio, and T. Raiko. Variational Bayes for continuous-time nonlinear state-space models. In *NIPS2006 Workshop on Dynamical Systems, Stochastic Processes and Bayesian Inference*, Whistler, B.C., Canada, 2006.

[20] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

[21] M. Harva and A. Kabán. Variational learning for rectified factor analysis. *Signal Processing*, 87(3):509–527, 2007.

[22] T. Raiko, A. Ilin and J. Karhunen. Principal component analysis for large scale problems with lots of missing values. In *Proc. of the 18th European Conf. on Machine Learning (ECML 2007)*, pages 691–698, Warsaw, Poland, September 2007.

[23] T. Raiko, A. Ilin, and J. Karhunen. Principal component analysis for sparse high-dimensional data. In *Proc. of the 14th Int. Conf. on Neural Information Processing (ICONIP 2007)*, Kitakyushu, Japan, November 2007.

[24] K. Kersting, L. De Raedt, and T. Raiko. Logical hidden Markov models. *Journal of Artificial Intelligence Research (JAIR)*, Vol. 25, pp. 425–456, April 2006.

[25] T. Raiko. Higher order statistics in play-out analysis. *Proc. of the Scandinavian Conf. on Artificial Intelligence, SCAI2006*, pp. 189–195, Espoo, Finland, October 25-27, 2006.

[26] M. Harva. A variational EM approach to predictive uncertainty. *Neural Networks*, 20(4):550–558, 2007.

[27] L. Nolan, M. Harva, A. Kabán, and S. Raychaudhury. A data-driven Bayesian approach for finding young stellar populations in early-type galaxies from their UV-optical spectra. *Monthly Notices of the Royal Astronomical Society*, 366(1):321–338, 2006.

[28] S. Refsdal. On the possibility of determining Hubble's parameter and the masses of galaxies from the gravitational lens effect. *Monthly Notices of the Royal Astronomical Society*, 128:307–310, 1964.

[29] J. C. Cuevas-Tello, P. Tino, and S. Raychaudhury. How accurate are the time delay estimates in gravitational lensing? *Astronomy & Astrophysics*, 454:695–706, 2006.

[30] M. Harva and S. Raychaudhury. Bayesian estimation of time delays between unevenly sampled signals. In *Proc. Int. Workshop on Machine Learning for Signal Processing (MLSP'06)*, pages 111–116. Maynooth, Ireland, 2006.

[31] M. Harva and S. Raychaudhury. Bayesian estimation of time delays between unevenly sampled signals. *Neurocomputing*, 2008. To appear.

Chapter 3

Independent component analysis and blind source separation

Erkki Oja, Juha Karhunen, Alexander Ilin, Antti Honkela, Karthikesh Raju, Tomas Ukkonen, Zhirong Yang, Zhijian Yuan

3.1 Introduction

What is Independent Component Analysis and Blind Source Separation? Independent Component Analysis (ICA) is a computational technique for revealing hidden factors that underlie sets of measurements or signals. ICA assumes a statistical model whereby the observed multivariate data, typically given as a large database of samples, are assumed to be linear or nonlinear mixtures of some unknown latent variables. The mixing coefficients are also unknown. The latent variables are nongaussian and mutually independent, and they are called the independent components of the observed data. By ICA, these independent components, also called sources or factors, can be found. Thus ICA can be seen as an extension to Principal Component Analysis and Factor Analysis. ICA is a much richer technique, however, capable of finding the sources when these classical methods fail completely.

In many cases, the measurements are given as a set of parallel signals or time series. Typical examples are mixtures of simultaneous sounds or human voices that have been picked up by several microphones, brain signal measurements from multiple EEG sensors, several radio signals arriving at a portable phone, or multiple parallel time series obtained from some industrial process. The term blind source separation is used to characterize this problem. Also other criteria than independence can be used for finding the sources.

Our contributions in ICA research. In our ICA research group, the research stems from some early work on on-line PCA, nonlinear PCA, and separation, that we were involved with in the 80's and early 90's. Since mid-90's, our ICA group grew considerably. This earlier work has been reported in the previous Triennial and Biennial reports of our laboratory from 1994 to 2005. A notable achievement from that period was the textbook "Independent Component Analysis" (Wiley, May 2001) by A. Hyvärinen, J. Karhunen, and E. Oja. It has been very well received in the research community; according to the latest publisher's report, over 5000 copies had been sold by August, 2007. The book has been extensively cited in the ICA literature and seems to have evolved into the standard text on the subject worldwide. In Google Scholar, the number of hits (in early 2008) is over 2300. In 2005, the Japanese translation of the book appeared (Tokyo Denki University Press), and in 2007, the Chinese translation (Publishing House of Electronics Industry).

Another tangible contribution has been the public domain FastICA software package (`http://www.cis.hut.fi/projects/ica/fastica/`). This is one of the few most popular ICA algorithms used by the practitioners and a standard benchmark in algorithmic comparisons in ICA literature.

In the reporting period 2006 - 2007, ICA/BSS research stayed as one of the core projects in the laboratory, with the pure ICA theory somewhat waning and being replaced by several new directions. Chapter 3 starts by introducing some theoretical advances on the FastICA algorithm undertaken during the reporting period, followed by a number of extensions of ICA and BSS. The first one is the method of independent subspaces with decoupled dynamics, that can be used to model complex dynamical phenomena. The second extension is related to Canonical Correlation Analysis, and the third one is nonnegative separation by the new Projective Nonnegative Matrix Factorization (P-NMF) principle. An application of ICA to telecommunications is also covered. Then the Denoising Source Separation (DSS) algorithm is applied to climate data analysis. This is an interesting and potentially very useful application that will be under intensive research in the future in the group.

Another way to formulate the BSS problem is Bayesian analysis. This is covered in the separate Chapter 2.

3.2 Convergence and finite-sample behaviour of the FastICA algorithm

Erkki Oja

In Independent Component Analysis, a set of original source signals are retrieved from their mixtures based on the assumption of their mutual statistical independence. The simplest case for ICA is the instantaneous linear noiseless mixing model. In this case, the mixing process can be expressed as

$$\mathbf{X} = \mathbf{AS}, \tag{3.1}$$

where $\mathbf{X}$ is a $d \times N$ data matrix. Its rows are the observed mixed signals, thus d is the number of mixed signals and N is their length or the number of samples in each signal. Similarly, the unknown $d \times N$ matrix $\mathbf{S}$ includes samples of the original source signals. $\mathbf{A}$ is an unknown regular $d \times d$ mixing matrix. It is assumed square because the number of mixtures and sources can always be made equal in this simple model.

In spite of the success of ICA in solving even large-scale real world problems, some theoretical questions remain partly open. One of the most central questions is the theoretical accuracy of the developed algorithms. Mostly the methods are compared through empirical studies, which may demonstrate the efficacy in various situations. However, the general validity cannot be proven like this. A natural question is, whether there exists some theoretical limit for separation performance, and whether it is possible to reach it.

Sometimes the algorithms can be shown to converge in theory to the correct solution giving the original sources, under the assumption that the sample size N is *infinite*. In [1], the FastICA algorithm was analyzed from this point of view. A central factor in the algorithm is a nonlinear function that is the gradient of the ICA cost function. It may be a polynomial, e.g. a cubic function in the case of kurtosis maximization/minimization, but it can be some other suitable nonlinearity as well. According to [1], let us present an example of convergence when the nonlinearity is the third power, and the 2×2 case is considered for the mixing matrix $\mathbf{A}$ in model (3.1).

In the theoretical analysis a linear transformation was made first, so that the correct solution for the separation matrix $\mathbf{W}$ (essentially the inverse of matrix $\mathbf{A}$) is a unit matrix or a variant (permutation and/or sign change). Thus the four matrix elements of $\mathbf{W}$ convergence to zero or to ± 1. The FastICA algorithm boils down to an iteration $w_{t+1} = f(w_t)$ for all the four elements of the separation matrix. The curve in Figure 3.2 shows the iteration function $f(.)$ governing this convergence. It is easy to see that close to the stable points, the convergence is very fast, because the iteration function is very flat.

In practice, however, the assumption of infinite sample size is unrealistic. For *finite* data sets, what typically happens is that the sources are not completely unmixed but some traces of the other sources remain in them even after the algorithm has converged. This means that the obtained demixing matrix $\widehat{\mathbf{W}}$ is not exactly the inverse of $\mathbf{A}$, and the matrix of estimated sources $\mathbf{Y} = \widehat{\mathbf{W}}\mathbf{X} = \widehat{\mathbf{W}}\mathbf{AS}$ is only approximately equal to $\mathbf{S}$. A natural measure of error is the deviation of the so-called gain matrix $\mathbf{G} = \widehat{\mathbf{W}}\mathbf{A}$ from the identity matrix, i.e., the variances of its elements.

The well-known lower limit for the variance of a parameter vector in estimation theory is the Cramér-Rao lower bound (CRB). In [2], the CRB for the demixing matrix of the FastICA algorithm was derived. The result depends on the score functions of the sources,

$$\psi_k(s) = -\frac{d}{ds}\log p_k(s) = -\frac{p_k'(s)}{p_k(s)} \tag{3.2}$$

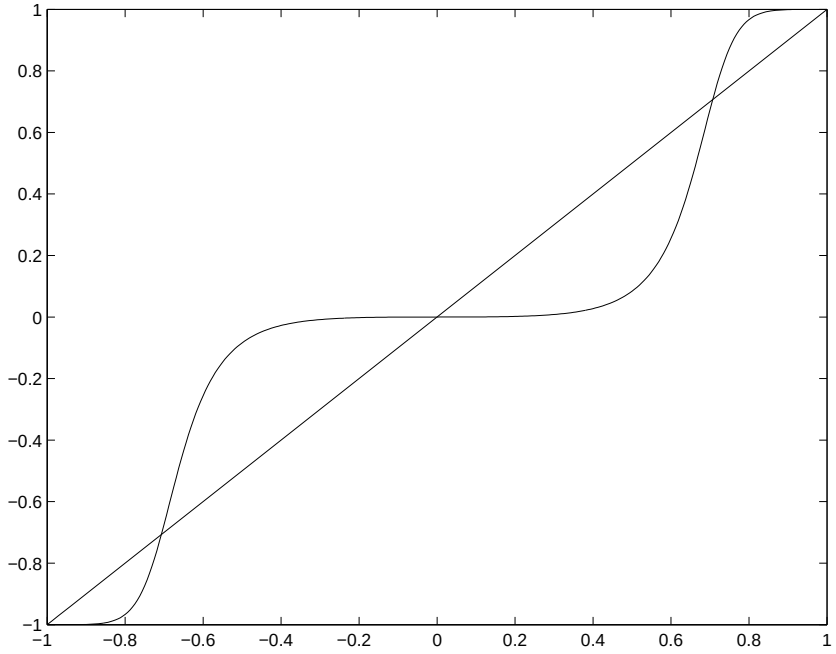


Figure 3.1: Shape of the iteration function for separation matrix elements, kurtosis case

where $p_k(s)$ is the probability density function of the k-th source. Let

$$\kappa_k = \mathrm{E}\left[\psi_k^2(s_k)\right]. \tag{3.3}$$

Then, assuming that the correct score function is used as the nonlinearity in the FastICA algorithm, the asymptotic variances of the off-diagonal elements (k, ℓ) of matrix $\mathbf{G}$ for the one-unit and symmetrical FastICA algorithm, respectively, read

$$V_{k\ell}^{1U-opt} = \frac{1}{N}\frac{1}{\kappa_k - 1} \tag{3.4}$$

$$V_{k\ell}^{SYM-opt} = \frac{1}{N}\frac{\kappa_k + \kappa_\ell - 2 + (\kappa_\ell - 1)^2}{(\kappa_k + \kappa_\ell - 2)^2}, \tag{3.5}$$

while the CRB reads

$$\mathrm{CRB}(\mathbf{G}_{k\ell}) = \frac{1}{N}\frac{\kappa_k}{\kappa_k\kappa_\ell - 1}. \tag{3.6}$$

Comparison of these results implies that the algorithm FastICA is nearly statistically efficient in two situations:

(1) One-unit version FastICA with the optimum nonlinearity is asymptotically efficient for $\kappa_k \to \infty$, regardless of the value of κ_ℓ.

(2) Symmetric FastICA is nearly efficient for κ_i lying in a neighborhood of 1^+, provided that all independent components have the same probability distribution function, and the nonlinearity is equal to the joint score function.

The work was continued to find a version of the FastICA that would be asymptotically efficient, i.e. able to attain the CRB. This can be achieved in the orthogonalization stage of the FastICA algorithm: instead of requiring strict orthogonalization, this condition is relaxed to allow small deviations from orthogonality, controlled by a set of free parameters. These parameters can be optimized so that the exact CRB is reached by the new algorithm, given that the correct score functions are used as nonlinearities.

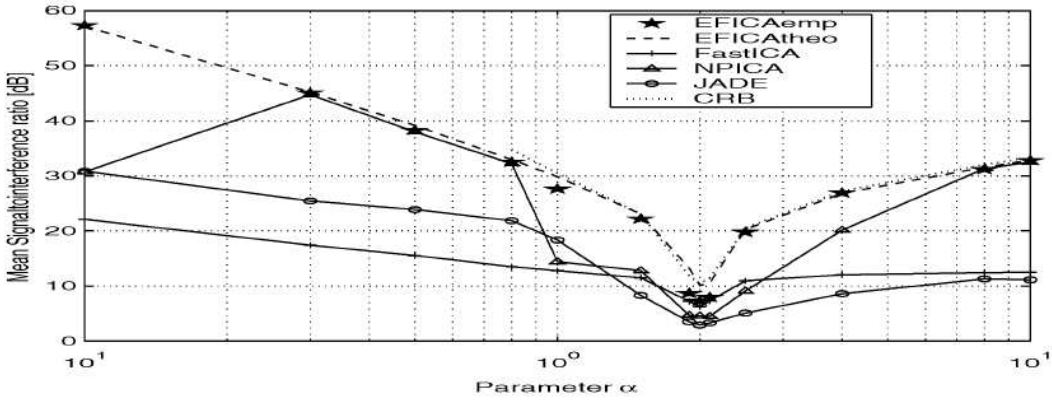


Figure 3.2: The Mean Signal-to Inference Ratio of EFICA, compared to CRB and some other ICA algorithms

The new efficient FastICA algorithm, dubbed EFICA, requires two phases because the score functions have to estimated first. Once they have been estimated, the new approximative orthogonalization scheme is run for a number of steps to reach the optimal solution. Figure 3.2 shows the efficiency of EFICA. To make meaningful comparisons, 13 source signals were artificially generated, each having a generalized gamma density $GG(\alpha)$ (where the value $\alpha = 2$ corresponds to the Gaussian density). The α values ranged from 0.1 to 10 and their places are marked by asterisks in the figure. The Mean Signal-to-Inference Ratio (SIR), both theoretical and experimental, obtained by EFICA is shown in the image (uppermost curve). It is very close to the Cramér-Rao Bound attainable in this situation, and far better than the Mean SIR attained by some other algorithms such as plain FastICA, NPICA, or JADE.

References

[1] Oja, E. and Yuan, Z.: The FastICA algorithm revisited – convergence analysis. *IEEE Trans. on Neural Networks 17*, no. 6, pp. 1370 - 1381 (2006).

[2] Tichavský, P., Koldovský, Z. and Oja, E.: Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis. *IEEE Trans. on Signal Processing 54*, no. 4, pp. 1189 - 1203 (2006).

[3] Koldovský, Z., Tichavský, P., and Oja, E.: Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound. *IEEE Trans. on Neural Networks 17*, no. 5, pp. 1265 - 1277 (2006).

3.3 Independent subspaces with decoupled dynamics

Alexander Ilin

Independent subspace models extend the general source separation problem by allowing groups (subspaces) $\mathbf{s}_k$ of sources:

$$\mathbf{x}(t) = \sum_{k=1}^{K} \mathbf{A}_k \mathbf{s}_k(t)\,. \tag{3.7}$$

The sources within one group $\mathbf{s}_k$ are assumed dependent while signals from different groups are mutually independent. Similarly to classical BSS, subspaces can be separated exploiting non-Gaussianity or temporal structures of the mixed signals. The technique presented in [2] uses a first-order nonlinear model to model the dynamics of each subspace:

$$\mathbf{s}_k(t) = \mathbf{g}_k(\mathbf{s}_k(t-1)) + \mathbf{m}_k(t)\,, \quad k = 1, \ldots, K\,, \tag{3.8}$$

Both the de-mixing transformation and the nonlinearities $\mathbf{g}_k$ governing the dynamics are estimated simulateously by minimizing the mean prediction error of the subspace dynamical models (3.8). The optimization procedure can be implemented using the algorithmic structure of denoising source separation [1].

The algorithm was tested on artificially generated data containing linear mixtures of two independent Lorenz processes with different parameters, a harmonic oscillator and two white Gaussian noise signals (see Fig. 3.3). The algorithm is able to separate the three subspaces using only the information about their dimensionalities.

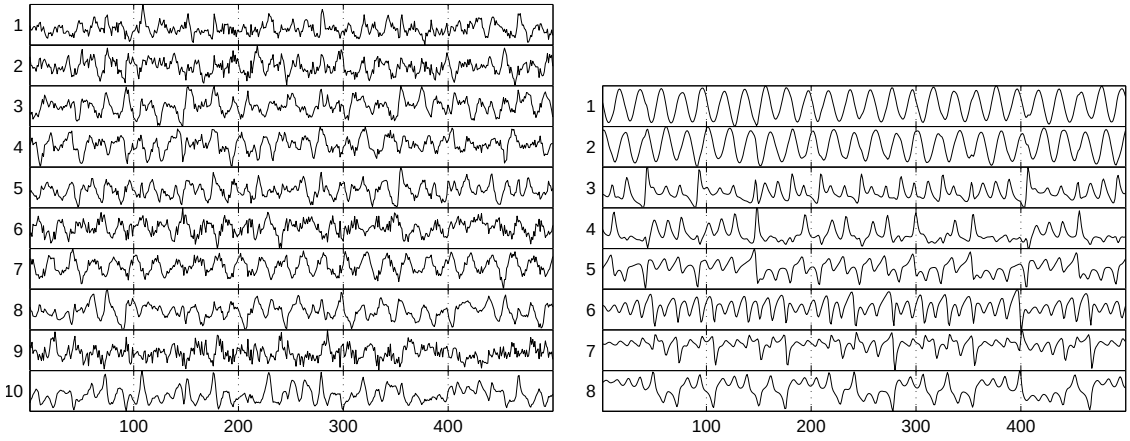


Figure 3.3: Left: Artificially generated linear mixtures of three dynamical processes and white noise signals. Right: Sources extracted by the technique extracting subspaces (signals 1–2, 3–5 and 6–9) with decoupled dynamics.

References

[1] J. Särelä and H. Valpola. Denoising source separation. *Journal of Machine Learning Research*, 6:233–272, 2005.

[2] A. Ilin. Independent dynamics subspace analysis. In *Proc. of the 14th European Symposium on Artificial Neural Networks (ESANN 2006)*, pp. 345–350, April 2006.

3.4 Extending ICA for two related data sets

Juha Karhunen, Tomas Ukkonen

Standard linear principal component analysis (PCA) [2, 1] and independent component analysis (ICA) [1] are both based on the same type of simple linear latent variable model for the observed data vector $\mathbf{x}(t)$:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{i=1}^{n} s_i(t)\mathbf{a}_i \tag{3.9}$$

In this model, the data vector $\mathbf{x}(t)$ is expressed as a linear transformation of the coefficient vector $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_n(t)]^T$. The column vectors $\mathbf{a}_i$, $i = 1, 2, \ldots, n$, of the transformation matrix $\mathbf{A}$ comprise the basis vectors of PCA or ICA, and the components $s_i(t)$ of the source vector $\mathbf{s}(t)$ are respectively principal or independent components corresponding to the data vector $\mathbf{x}(t)$. For simplicity, we assume that both the data vector $\mathbf{x}(t)$ and the source vector $\mathbf{s}(t)$ are zero mean n-vectors, and that the basis matrix $\mathbf{A}$ is a full-rank constant $n \times n$ matrix.

In PCA, the basis vectors $\mathbf{a}_i$ are required to be mutually orthogonal, and the coefficients $s_i(t)$ to have maximal variances (power) in the expansion (3.9) [2, 1]. While in ICA the basis vectors $\mathbf{a}_i$ are generally non-orthogonal, and the expansion (3.9) is determined under certain ambiguities from the strong but often meaningful condition that the coefficients $s_i(t)$ must be mutually statistically independent or as independent as possible [1].

Canonical correlation analysis (CCA) [2] is a generalization of PCA for two data sets whose data vectors are denoted by $\mathbf{x}$ and $\mathbf{y}$. CCA seeks for the linear combinations of the components of the vectors $\mathbf{x}$ and $\mathbf{y}$ which are maximally correlated. In this work, we have considered a similar expansion as (3.9) for both $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad \mathbf{y} = \mathbf{B}\mathbf{t} \tag{3.10}$$

We then try to find in a similar manner as in ICA the maximally independent and dependent components from $\mathbf{x}$ and $\mathbf{y}$ by using higher-order statistics. As a result, we get an ICA style counterpart for canonical correlation analysis.

These ideas are introduced in [3], and discussed in more detail in the journal paper [4]. The methods introduced in these papers are somewhat heuristic, buut seem to work adequately both for artificially generated data and in a difficult cryptographic problem. We also consider in these papers practical measures for statistical dependence or independence of two random variables.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001, 481+xxii pages.

[2] A. Rencher, *Methods of Multivariate Analysis, 2nd ed.* Wiley, 2002.

[3] J. Karhunen and T. Ukkonen. Generalizing independent component analysis for two related data sets. In *Proc. of the IEEE 2006 Int. Conf. on Neural Networks / 2006 IEEE World Congress on Computational Intelligence (IJCNN2006/WCCI2006)*, Vancouver, Canada, July 2006, pp. 1822-1829.

[4] J. Karhunen and T. Ukkonen, Extending ICA for finding jointly dependent components from two related data sets. *Neurocomputing*, Vol. 70, Issues 16-18, October 2007, pp. 2969-2769.

3.5 ICA in CDMA communications

Karthikesh Raju, Tapani Ristaniemi, Juha Karhunen, Erkki Oja

In wireless communication systems, like mobile phones, an essential issue is division of the common transmission medium among several users. A primary goal is to enable each user of the system to communicate reliably despite the fact that the other users occupy the same resources, possibly simultaneously. As the number of users in the system grows, it becomes necessary to use the common resources as efficiently as possible.

During the last years, various systems based on CDMA (Code Division Multiple Access) techniques [1, 2] have become popular, because they offer several advantages over the more traditional FDMA and TDMA schemes based on the use of non-overlapping frequency or time slots assigned to each user. Their capacity is larger, and it degrades gradually with increasing number of simultaneous users who can be asynchronous. On the other hand, CDMA systems require more advanced signal processing methods, and correct reception of CDMA signals is more difficult because of several disturbing phenomena [1, 2] such as multipath propagation, possibly fading channels, various types of interferences, time delays, and different powers of users.

Direct sequence CDMA data model can be cast in the form of a linear independent component analysis (ICA) or blind source separation (BSS) data model [3]. However, the situation is not completely blind, because there is some prior information available. In particular, the transmitted symbols have a finite number of possible values, and the spreading code of the desired user is known.

In this project, we have applied independent component analysis and denoising source separation (DSS) to blind suppression of various interfering signals appearing in direct sequence CDMA communication systems. The standard choice in communications for suppressing such interfering signals is the well-known RAKE detection method [2]. RAKE utilizes available prior information, but it does not take into account the statistical independence of the interfering and desired signal. On the other hand, ICA utilizes this independence, but it does not make use of the prior information. Hence it is advisable to combine the ICA and RAKE methods for improving the quality of interference cancellation.

In the journal paper [4], various schemes combining ICA and RAKE are introduced and studied for different types of interfering jammer signals under different scenarios. By using ICA as a preprocessing tool before applying the conventional RAKE detector, some improvement in the performance is achieved, depending on the signal-to-interference ratio, signal-to-noise ratio, and other conditions [4].

All these ICA-RAKE detection methods use the FastICA algorithm [3] for separating the interfering jammer signal and the desired signal. In the case of multipath propagation, it is meaningful to examine other temporal separation methods, too. We have also applied denoising source separation [5] to interference cancellation. This is a semi-blind approach which uses the spreading code of the desired user but does not require training sequences. The results of the DSS-based interference cancellation scheme show improvements over conventional detection.

All the results achieved in this project have been collected and presented in the monograph type doctoral thesis [6].

References

[1] S. Verdu, *Multiuser Detection*. Cambridge Univ. Press, 1998.

[2] J. Proakis, *Digital Communications*. McGraw-Hill, 3rd edition, 1995.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001, 481+xxii pages.

[4] K. Raju, T. Ristaniemi, J. Karhunen, and E. Oja, Jammer cancellation in DS-CDMA arrays using independent component analysis. *IEEE Trans. on Wireless Communications*, Vol. 5, No. 1, January 2006, pp. 77–82.

[5] J. Särelä and H. Valpola, Denoising source separation. *J. of Machine Learning Research*, Vol. 6, 2005, pp. 233–272.

[6] K. Raju, *Blind Source Separation for Interference Cancellation in CDMA Systems.* PhD Thesis, Helsinki Univ. of Technology, 2006. Published as Report D16, Laboratory of Computer and Information Science.

3.6 Non-negative projections

Zhirong Yang, Jorma Laaksonen, Zhijian Yuan, Erkki Oja

Projecting high-dimensional input data into a lower-dimensional subspace is a fundamental research topic in signal processing, machine learning and pattern recognition. Non-negative projections are desirable in many real-world applications where the original data are non-negative, consisting for example of digital images or various spectra. It was pointed out by Lee and Seung [1] that the positivity or non-negativity of a linear expansion is a very powerful constraint, that seems to lead to sparse representations for the data. Their method, *non-negative matrix factorization (NMF)*, minimizes the difference between the data matrix $\mathbf{X}$ and its non-negative decomposition $\mathbf{WH}$. The difference can be measured by the Frobenius matrix norm or the Kullback-Leibler divergence.

Yuan and Oja [2] proposed the *projective non-negative matrix factorization (P-NMF)* method which replaces $\mathbf{H}$ in NMF with $\mathbf{W}^T\mathbf{X}$. This actually combines the objective of principal component analysis (PCA) with the non-negativity constraint. The P-NMF algorithm has been applied to facial image processing [4] using a popular database, FERET [3]. Figure (3.4) visualizes the basis images learned by NMF and P-NMF. The empirical results indicate that P-NMF is able to produce more spatially localized, part-based representations of visual patterns.

Another attractive feature of the NMF and P-NMF methods is that their multiplicative update rules do not involve human-specified parameters such as the learning rate. Thus the analysis results are completely data driven. In [5] we have studied how to construct multiplicative update rules for non-negative projections based on Oja's iterative learning rule. Our method integrates the multiplicative normalization factor into the original additive update rule as an additional term which generally has a roughly opposite direction. As a consequence, the modified additive learning rule can easily be converted to its multiplicative version, which maintains the non-negativity after each iteration. With this technique, almost identical results to P-NMF can be obtained by imposing the non-negativity constraint on linear Hebbian networks.

The derivation of our approach provides a sound interpretation of learning non-negative projection matrices based on iterative multiplicative updates—a kind of Hebbian learning with normalization. A convergence analysis is provided by interpreting the multiplicative

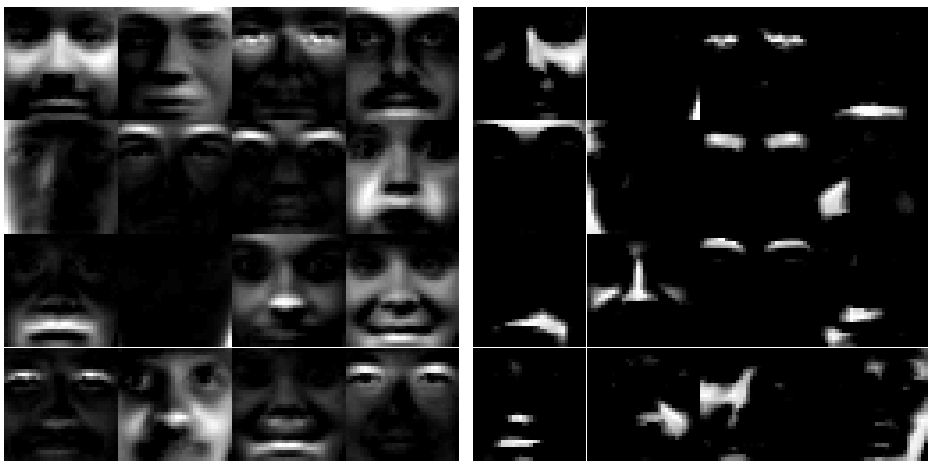


Figure 3.4: NMF (left) and P-NMF (right) bases of 16 dimensions.

updates as a special case of natural gradient learning. Furthermore, our non-negative variant of *linear discriminant analysis (LDA)* can serve as a feature selector. Its kernel extension can reveal an underlying factor in the data and be used as a sample selector.

References

[1] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[2] Zhijian Yuan and Erkki Oja. Projective nonnegative matrix factorization for image compression and feature extraction. In *Proc. of 14th Scandinavian Conference on Image Analysis (SCIA 2005)*, pages 333–342, Joensuu, Finland, June 2005.

[3] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1090–1104, October 2000.

[4] Zhirong Yang, Zhijian Yuan, and Jorma Laaksonen. Projective non-negative matrix factorization with applications to facial image processing. *International Journal on Pattern Recognition and Artificial Intelligence*, 21(8):1353–1362, December 2007.

[5] Zhirong Yang and Jorma Laaksonen. Multiplicative updates for non-negative projections. *Neurocomputing*, 71(1-3):363–373, 2007.

3.7 Climate data analysis with DSS

Alexander Ilin, Harri Valpola, Erkki Oja

An important task for which statistical methods are used in climate research is seeking physically meaningful interpretations of observed climate variability, for example, identification of 'modes' in the observational record. Statistical techniques which are widely used in this task include principal component analysis (PCA) or empirical orthogonal functions (EOFs), extended EOFs, and Hilbert EOFs [1]. Although EOFs have probably been the most popular tool for an efficient representation of climate records, EOF representation may be intuitively meaningless in a meteorological sense [2]. Therefore several techniques of rotated PCA/EOF have been proposed to ensure easier interpretation of the results. The rotation is realized using a linear transformation of principal components such that a suitably chosen criterion of "simple structure" is optimized. The objective is to find a data representation allowing for compact scientific explanation of a variable with a smaller number of principal components. Different assumptions on simplicity yield different rotation techniques.

We extend the concept of rotated PCA by introducing the concept of "interesting structure". In our case, the goal of exploratory analysis is to find signals with some specific structures of interest. They may for example manifest themselves mostly in specific variables, which exhibit prominent variability in a specific timescale etc. An example of such analysis can be extracting clear trends or quasi-oscillations from climate records. The procedure for obtaining suitable rotations of EOFs can be based on the general algorithmic structure of denoising source separation (DSS) [3].

In our initial studies, we tested the effectiveness of the proposed methodology to discover climate phenomena which are well-known in climatology, using very little information about their properties. One of the most prominent results is the extraction of the El Niño–Southern Oscillation phenomenon, using only a very generic assumption of its prominent variability in the interannual timescale (see Figs. 3.5-3.6) [4]. Other prominent signals found in this analysis might correspond to significant climate phenomena as well; for example, the second signal with prominent interannual variability somewhat resembles the derivative of the El Niño index (see Fig. 3.5).

Several other techniques for studying prominent climate variations have been introduced in our papers [4, 5]. Analysis which separates prominent quasi-oscillations in climate records by their frequency contents gives a meaningful representation of the slow climate variability as combination of trends, interannual oscillations, the annual cycle and slowly changing seasonal variations [4]. The technique presented in [5] can be used for studying slow variability present in fast weather fluctuations.

The results of the climate research were presented at the Fifth Conference on Artificial Intelligence Applications to Environmental Science as part of the 87th Annual Meeting of the American Meteorological Society (best student presentation) [6] and at the 10th International Meeting on Statistical Climatology.

References

[1] H. von Storch, and W. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge, U.K, 1999.

[2] M. B. Richman. Rotation of principal components. *Journal of Climatology*, 6:293–335, 1986.

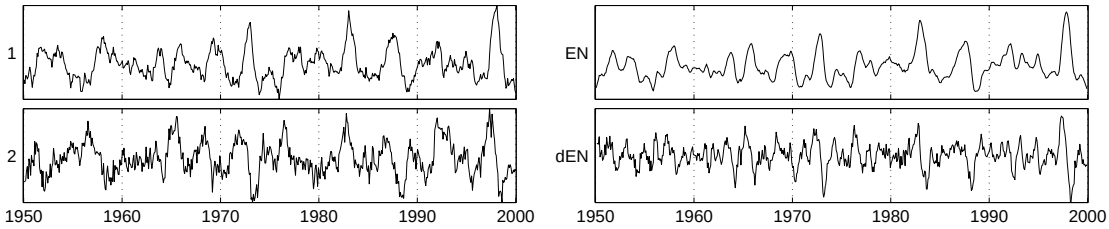


Figure 3.5: Left: The time courses of the two interannual phenomena found in global temperature, air pressure and precipitation data using DSS. Right: The index used in climatology to measure the strength of El Niño (marked as EN) and the derivative of the El Niño index (marked as dEN). The similarity is striking for the upper signals and some common features can be observed in the lower signals.

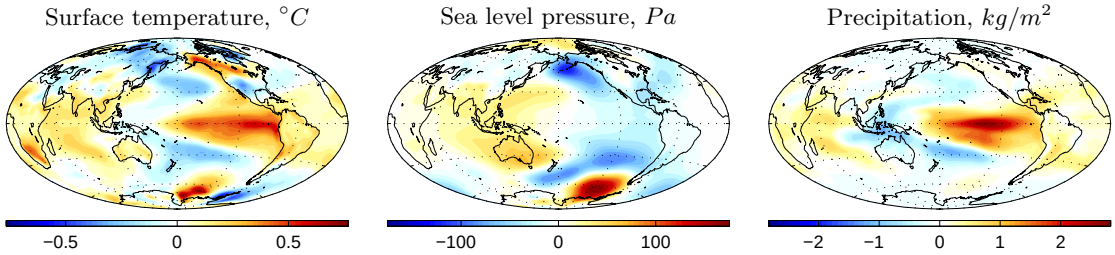


Figure 3.6: Spatial patterns corresponding to the most prominent interannual phenomenon found in climate data. The maps display the regions in which the effect of the phenomenon is most prominent. The maps contain many features traditionally associated with El Niño–Southern Oscillation phenomenon.

[3] J. Särelä and H. Valpola. Denoising source separation. *Journal of Machine Learning Research*, 6:233–272, 2005.

[4] A. Ilin, and H. Valpola, and E. Oja. Exploratory analysis of climate data using source separation methods. *Neural Networks*, Vol. 19, No. 2, pp. 155–167, March 2006.

[5] A. Ilin, and H. Valpola, and E. Oja. Extraction of components with structured variance. In *Proc. of the IEEE World Congress on Computational Intelligence (WCCI 2006)*, pp. 10528–10535, Vancouver, BC, Canada, July 2006.

[6] A. Ilin, and H. Valpola, and E. Oja. Finding interesting climate phenomena by exploratory statistical techniques. In *Proc. of the Fifth Conference on Artificial Intelligence Applications to Environmental Science as part of the 87th Annual Meeting of the American Meteorological Society*, San Antonio, TX, USA, January 2007. Best student presentation.

Chapter 4

Modeling of relevance

Samuel Kaski, Jaakko Peltonen, Kai Puolamäki, Janne Sinkkonen, Jarkko Venna, Arto Klami, Jarkko Salojärvi, Eerika Savia

4.1 Introduction

We develop statistical machine learning methods for extracting useful regularities from large, high-dimensional data sets. The key concept is *modeling of relevance*: data are usually full of patterns but the extracted ones should obviously be relevant to the analyst. An explicit definition of what is relevant is usually not known, and relevance needs to be inferred indirectly.

We have developed methods that use the structure of data in constraining which kinds of regularities are considered relevant. The structure here means several data sources or data sets. To make the task more concere, we have divided the ways of using the structure of data into three subtypes:

- *Relevance through data fusion* can mean two principal things: *dependency mining* and *supervised mining*, which are applicable in different settings. In both, several sources are combined with the goal of identifying relevant *aspects*, features or feature combinations, of data.

  In *dependency mining* or exploration, the aim is to decompose variation in each data source into source-specific and shared components. The within-source variation is assumed irrelevant, "noise", and only the shared effects are relevant. An example is measurement of several noisy signals from a common source, when characteristics of the noise are not known. More examples are given in Sections 5 and 6.

  While dependency mining is symmetric, in *supervised mining* a supervising auxiliary data set supervises the mining of primary data. Otherwise the methods are similar. If the supervising set consists of class labels of the primary data samples, the setup is *supervised unsupervised learning*. Our earlier research topic *learning metrics* was one suitable method for supervised unsupervised learning.

- *Relevant subtask learning* is a new research topic we introduced for addressing the problem of having too little representative or known-to-be-relevant training data. Given that other, partly or wholly irrelevant data sets are available, the relevant small data set is used as a "query" to retrieve more relevant data. At the same time, a model is built using all relevant data.

  This work can be seen as a special kind of asymmetric multi-task learning, or as combining information retrieval with multi-task learning.

- For *modeling of networks* we develop scalable models capable of dealing with uncertainty in network data. Networks are the simplest kinds of relational data, where the relations give hints of relevance.

These two general topics are useful in most of the modeling tasks above:

- *Discriminative generative modeling* describes how to use rigorous statistical modeling machinery for learning what is relevant to classes, and for making inference.

- *Information visualization* is a central subproblem in exploratory analysis and mining. We have introduced new very competitive nonlinear projection methods particularly suitable for projection to small dimensions for visualization.

4.2 Relevance through data fusion

Unsupervised data exploration or mining is defined as search for systematic properties, statistical structures or patterns from data. The findings need to be *relevant* as well, and typically relevance has been defined implicitly by selecting which kinds of patterns to find, which distance measures or features to use, and which model family to use. In general, relevance is defined by bringing in prior knowledge or assumptions to the task.

We have introduced methods for bringing in the prior information in a data-driven way, by choosing additional data sources and defining relevance through statistical dependencies between the sources. The underlying assumption is that aspects of data that are visible in one source only are "noise", whereas aspects visible in several sources describe the common thing of which all sources have different views. This will become clearer in the detailed descriptions below.

A straightforward way of finding the shared view is to build representations of data from each source, by maximizing the statistical dependency of the representations of different sources. We have developed both theory and practical methods for this task, and applied the methods in particular in neuro- and bioinformatics (Chapters 6 and 5).

Probabilistic models for detecting dependencies

Above the general approach was formulated in terms of maximizing a chosen dependency measure for mappings, that is, representations of the observations. We have previously introduced various methods for this task, including *associative clustering* and a linear projection method maximizing a non-parametric estimate of mutual information.

Recently we have studied an alternative formulation for the same task. One of the central problems in data analysis is overlearning, which means that models estimated with small data sets do not generalize well to new observations. One common solution to overlearning is to apply *Bayesian analysis* that allows treating uncertainties and choosing model complexity in a justified manner. Prior information can be rigorously incorporated to improve learning from small data sets, it is straightforward to extend models by changing distributional assumptions, and it is easy to construct larger models by combining submodels, at least in principle.

Bayesian tools can be applied to probabilistic models providing a generative description of the observed data. The methods for detecting dependencies between data sets are not, however, formulated as such models, and hence the Bayesian approach has not been possible for this task. We have introduced new theory on how such models can be built [3], and presented example models derived from the theory.

The proposed model family consists of latent variable models (see Fig. 4.1), where the observed data of each source is assumed to be an additive composition of two sources: one that is shared with the other data sources, and one that is specific to that particular data source. We have shown [3] that such models can extract the statistical dependencies to the shared latent source if a particular requirement is satisfied: the part of the model describing the data-source-specific variation in the observed data should be accurate enough.

Based on this basic principle we have re-derived an earlier probabilistic interpretation of canonical correlation analysis [1], and provided two novel models. In [3] a Bayesian clustering model for detecting dependencies is solved with variational Bayes approximation. The model is illustrated graphically in Figure 4.1. In [2] a Bayesian version of canonical correlation analysis is introduced, this time using Gibbs sampling for inference. Besides introducing a way of analyzing small-sample data to CCA the method lifts a critical restriction of classical CCA: the requirement of global linear dependency. It is overcome by

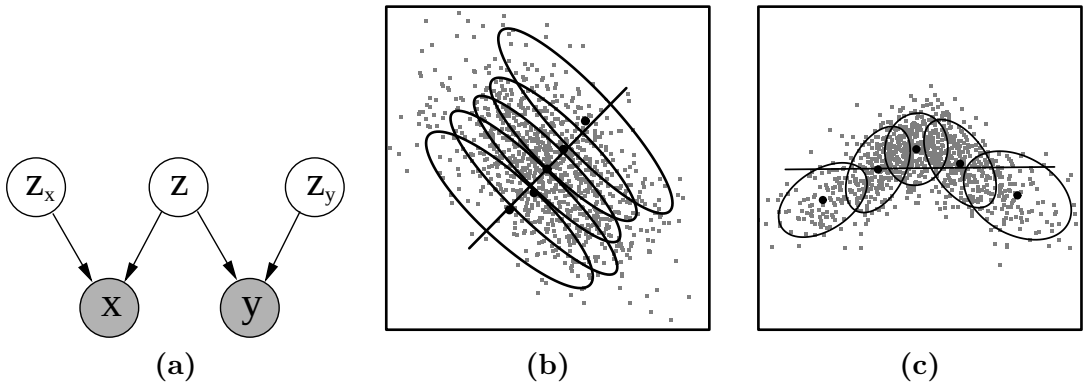


Figure 4.1: **(a):** A general latent-variable model structure for detecting statistical dependencies between **x** and **y**. **(b-c):** Illustration of a clustering version of the general model. The two panels show scatter-plots of two data sets having co-occurring samples. The lines depict linear dependency found by canonical correlation analysis. The clusters found by the clustering model have aligned according to the dependency, while still capturing nonlinear structure of the data in panel **(c)**.

introducing a Dirichlet process mixture model, allowing different kinds of dependencies in different parts of the data space.

Dependency with class variables

A common case of two data sources is class labels coupled with feature vectors. Standard classifiers use the dependencies between the sources to predict class labels; other applications include visualization, discriminative clustering or discriminative feature extraction. These tasks use the labels to guide unsupervised analysis of the features; we call them *supervised unsupervised learning*.

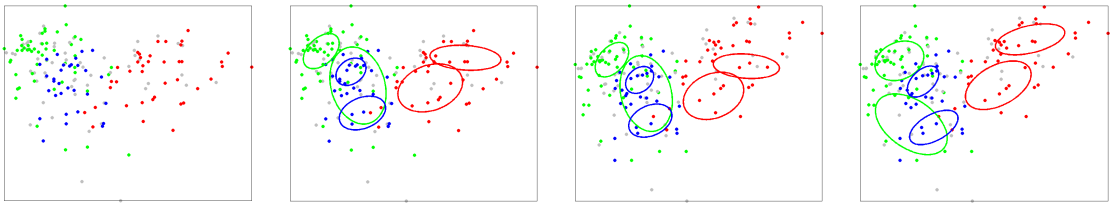
Recently we have studied a particular application of supervised unsupervised learning: fast learning of a class-discriminative subspace of data features. The subspace is defined by a linear transformation, and the features in the class-discriminative subspace are *discriminative components* of data. The subspace is useful for visualization, dimensionality reduction, feature extraction, and for learning a regularized distance metric.

Earlier we had learned such transformations with nonparametric estimation [5] which is accurate but slow; the computational complexity is $O(N^2)$ per iteration; here N is the number of samples. We now introduced a method that learns the linear transformation in a fast, semisupervised way [4], by optimizing a mixture model for classes in the subspace. The new method (Fig. 4.2) is fast ($O(N)$ per iteration) and semi-supervised, that is, can use unlabeled and pairwise-constrained data as well as labeled data.

References

[1] Fransic R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. Tech. Rep 688, Department of Statistics, University of California, Berkeley, 2005.

[2] Arto Klami and Samuel Kaski. Local dependent components. In Zoubin Ghahramani, editor, *Proceedings of the 24th International Conference on Machine Learning*, pages 425–432, 2007.



(a) (b) (c) (d)

Figure 4.2: Sample iterations of optimizing the discriminative subspace. Dots show data in the subspace; ellipses show the shape of mixture model components used to model the distribution in the subspace. There are three classes (red, green, blue) and unlabeled samples (gray dots). **(a):** Initial transformation. **(b):** The mixture model is optimized for the transformation. **(c):** The transformation is optimized for the mixture model. **(d):** The mixture model is optimized for the new transformation. The iteration continues in alternating steps.

[3] Arto Klami and Samuel Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, accepted for publication, 2008.

[4] Jaakko Peltonen, Jacob Goldberger, and Samuel Kaski. Fast Semi-supervised Discriminative Component Analysis. In Konstantinos Diamantaras, Tülay Adali, Ioannis Pitas, Jan Larsen, Theophilos Papadimitriou, and Scott Douglas, editors, *Machine Learning for Signal Processing XVII*, pages 312–317. IEEE, 2007.

[5] Jaakko Peltonen and Samuel Kaski. Discriminative components of data. *IEEE Transactions on Neural Networks*, 16: 68–83, 2005.

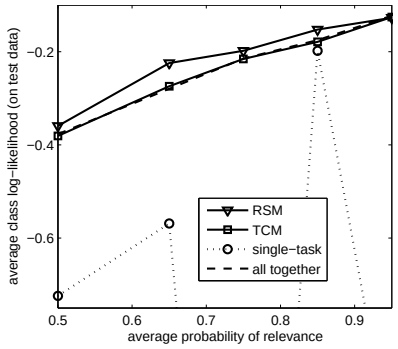


Figure 4.3: Relevant subtask learning model (RSM) outperforms a multi-task method that clusters tasks (TCM) and to two naive methods ("single-task" and "all together"), on news article data. The task was to predict relevance of news articles to a specific reader (the reader-of-interest), using articles rated by other readers as additional sources of information. Average results over 10 generated problems are shown, as a function of one experiment design parameter, the average probability that a sample is relevant to the reader-of-interest.

4.3 Relevant subtask learning

Having too little labeled training data is a common problem in classifier design. The problem is particularly hard for the high-dimensional data in genome-wide studies of modern bioinformatics, but appears also in image classification from few examples, finding of relevant texts, etc.

After realizing that the world is full of other data sets, the problem becomes how to simultaneously learn from a small data set and retrieve useful information from the other data sets. We have recently introduced a learning problem called *relevant subtask learning*, a variant of multi-task learning, which aims to solve the small-data problem by intelligently making use of other, potentially related "background" data sets.

Such potentially related "background" data sets are available for instance in bioinformatics, where there are databases full of data measured for different tasks, conditions or contexts; for texts there is the web. Such data sets are *partially relevant*: they do not come from the exact same distribution as future test data, but their distributions may still contain some useful part. Our research problem is, *can we use the partially relevant data sets to build a better classifier for the test data?*

Learning from one of the data sets is called a "task". Our scenario is then a special kind of *multi-task learning* problem. However, in contrast to typical multi-task learning, our problem is fundamentally asymmetric and more structured; test data fits one task, the "*task-of-interest*," and other tasks may contain *subtasks* relevant for the task-of-interest, but no other task needs to be wholly relevant.

In [1] we introduced a method that uses logistic regression classifiers. The key is to assume that each data set is a mixture of relevant and irrelevant samples. By fitting this model to all data sets, the common model for relevant samples learns from all tasks. We model the irrelevant part with a sufficiently flexible model such that irrelevant samples cannot distort the model for relevant data. A sample application is a news recommender for one user, where classifications from other users are available (Fig. 4.3). The relevant subtask learner outperforms a comparable standard multi-task learning model (related to [2]).

References

[1] Samuel Kaski and Jaakko Peltonen. Learning from relevant tasks only. In Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, *Machine Learning: ECML 2007*, pages 608–615. Springer-Verlag, Berlin, Germany, 2007.

[2] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-Task Learning for Classification with Dirichlet Process Priors. *Journal of Machine Learning Research*, 8: 35–63, 2007.

4.4 Discriminative generative modeling

The more traditional counterpart to supervised mining is *discriminative learning* where the data set is the same but the task is different. Given paired data $(\mathbf{x}, c)$, the task is to predict c for a test set where only the values of $\mathbf{x}$ are known.

There exist two traditional modeling approaches for predicting c, discriminative and generative. Discriminative models optimize the conditional probability $p(c|\mathbf{x})$ (or some other discriminative criterion) directly. The models are good classifiers since they do not waste resources on modeling those properties of the data that do not affect the value of c, that is, the marginal distribution of $\mathbf{x}$. The alternative approach is generative modeling of the joint distribution $p(c, \mathbf{x})$. Generative models add prior knowledge of the distribution of $\mathbf{x}$ into the task. This facilitates for example inferring missing values, since the model is assumed to generate also the covariates $\mathbf{x}$. The generative models are often additionally simpler to construct, and their parameters offer simple explanations in terms of expected sufficient statistics.

Discriminative Joint Density Models. In discriminative generative modeling we study discriminative inference given a generative model family $p(c, \mathbf{x}, \theta)$. The model family is assumed to be as good as possible but still known to be incorrect, and the objective is to obtain a distribution or point estimate that is optimal for predicting the values of c given $\mathbf{x}$. The Bayesian approach of using the posterior of the generative model family $p(c, \mathbf{x}, \theta)$ is not particularly well justified in this case, and it is known that it does not always generalize well to new data [1].

One way of learning discriminative classifiers is to take a joint density model, and then change the objective function from joint likelihood $\prod_i p(c_i, \mathbf{x}_i|\theta)$ to conditional likelihood $\prod_i p(c_i|\mathbf{x}_i, \theta)$. Earlier, we have presented an EM algorithm for obtaining discriminative point estimates [2]. The point estimate is (asymptotically) consistent for discrimination, given the model family. In [3] we proved that this applies for distributions as well; we derived an axiomatic proof that a *discriminative posterior* is consistent for conditional inference; using the discriminative posterior is standard practice in Bayesian regression, but we show that it is rigorous for model families of joint densities as well.

Compared to pure discriminative models, the benefit of the approach is that prior knowledge about $\mathbf{x}$ is brought in. The models operate in the same parameter space as ordinary discriminative models, but the generative formulation constrains the model manifold. Additionally, the density estimate for $\mathbf{x}$ from the model can be used for inferring missing values in the data [3].

References

[1] Peter D. Grünwald and John Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66(2–3):119–149, 2007.

[2] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. Expectation maximization algorithms for conditional likelihoods. In Luc De Raedt and Stefan Wrobel, editors, *Proceedings of the 22nd International Conference on Machine Learning (ICML-2005)*, pages 753–760, New York, USA, 2005. ACM press.

[3] Kai Puolamäki, Jarkko Salojärvi, Eerika Savia, and Samuel Kaski. Discriminative MCMC. Report E1, Publications in Computer and Information Science, Helsinki University of Technology, 2006.

4.5 Visualization methods

Visualization of mutual similarities of entries in large high-dimensional data sets is a central subproblem in exploratory analysis and mining. It is makes sense to "look at the data" in all stages of data analysis, and reducing the dimensionality to two or three gives a scatterplot visualization.

When the intrinsic dimensionality of the data is higher than the dimensionality of the visualization, as is often the case, the visualization cannot represent the data flawlessly; some properties are necessarily lost or misrepresented. A compromise is unavoidable, but which compromise is the best for visualization? Many existing nonlinear dimensionality reduction methods practically ignore this question altogether, because they are not designed to reduce the dimensionality of the data set lower than is possible without losing information. Some methods choose the compromise implicitly in that they produce the lower-dimensional representation by minimizing a cost function, but the cost function has not been motivated from the point of view of visualization, that is, it is not obvious why a projection that minimizes the cost function should be a good visualization. We have filled this gap by introducing rigorously motivated measures for the quality of a visualization, as well as a nonlinear dimensionality reduction method that optimizes these measures and is therefore specifically designed for optimal visualization.

Visualization as information retrieval

We view visualization as an information retrieval task. Consider an analyst studying a scatterplot of countries, organized according to their welfare indicators. Being interested in Finland, she wants to know which other countries are similar. The visualization helps in this task of retrieving similar items, and quality of retrieval can be measured with standard information retrieval measures *precision* and *recall*. Any information retrieval method needs to make a compromise between these measures, parameterized by the relative cost of false positives and misses. Since a visualizer is an information retrieval device as well, it needs to make the same compromise.

We have adapted the information retrieval measures to visualization by smoothing them and representing them as differences between distributions of points being neighbors. It turns out that the traditional measures are limiting cases of these more general measures. Once the relative cost λ of false positives and misses has been fixed, we can directly optimize the visualization to minimize the retrieval cost. We call the resulting visualization method the Neighborhood Retrieval Visualizer (NeRV) [1].

The NeRV is a further development of our earlier method *local multidimensional scaling* [2], a faster method where the trade-off between precision and recall was heuristic and hence the results were less accurate.

Later we added the Self-Organizing Map to the comparison [3]. The SOM was very good in terms of (smoothed) precision, even producing a slightly better result than NeRV in some cases. In terms of recall the SOM performed poorly.

References

[1] Jarkko Venna and Samuel Kaski. Nonlinear dimensionality reduction as information retrieval. In Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS*07), San Juan, Puerto Rico, March 21-24, 2007.

[2] Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19:889–899, 2006.

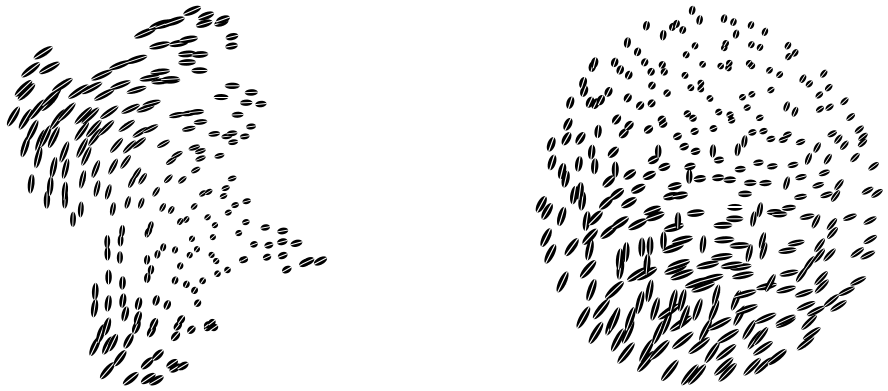


Figure 4.4: Two nonlinear projections of data that lies on the surface of a three-dimensional sphere. One of the input coordinates governs the rotation of the glyphs, the second their scale, and the third their degree of elongation. Hence, points having similar glyphs are close to each other in the input space. On the *left*, precision has been maximized; the sphere has become split open and the glyphs change smoothly, but on the opposite ends of the projection there are similar glyphs that are projected far from each other. On the *right*, recall has been maximized and the sphere has become squashed flat. There are areas where the different kinds of glyphs are close to each other, but there are no areas where similar glyphs are very far from each other.

[3] Kristian Nybo, Jarkko Venna and Samuel Kaski. The Self-Organizing Map as a Visual Neighbor Retrieval Method. In *Proceedings of 6th Int. Workshop on Self-Organizing Maps (WSOM '07)*. Bielefeld University, Bielefeld, Germany, 2007.

4.6 Networks

Machine Learning is in the midst of a "structural data revolution". After many decades of focusing on independent and identically-distributed examples, many researchers are now modelling inter-related entities that are linked together into complex graphs. A major driving force is the explosive growth of heterogeneous data collected on diverse sectors of the society. Example domains include bioinformatics, communication networks, and social network analysis.

Networks are a special case of structural data. Inferring properties of the network nodes, or vertices, from the links, or edges, has become a common data mining problem. Network data are typically not a complete description of reality but come with errors, omissions and uncertainties. Some links may be spurious, for instance due to measurement noise in biological networks, and some potential links may be missing, for instance friendship links of newcomers in social networks. Probabilistic generative models are a tool for modeling and inference under such uncertainty. They treat the links as random events, and give an explicit structure for the observed data and its uncertainty. Compared to non-stochastic methods, they are therefore likely to perform well as long as their assumptions are valid; they may reveal properties of networks that are difficult to observe with non-statistical techniques from the noisy and incomplete data, and they also offer a groundwork for new conceptual developments.

Component models for large networks

Being among the easiest ways to find meaningful structure from discrete data, Latent Dirichlet Allocation (LDA) and related component models have been applied widely. They are simple, computationally fast and scalable, interpretable, and admit flexible nonparametric priors. In the currently popular field of network modeling, relatively little work has taken uncertainty of data seriously in the Bayesian sense, and component models have been introduced to the field only recently. We have developed a component model of networks that finds community-like structures like the earlier methods motivated by physics. With Dirichlet Process priors and an efficient implementation the models are highly scalable.

References

[1] Janne Sinkkonen, Janne Aukia, and Samuel Kaski. Inferring vertex properties from topology in large networks. In *The 5th International Workshop on Mining and Learning with Graphs (MLG'07)*, Florence, Italy, 2007. Universita Degli Studi di Firenze.

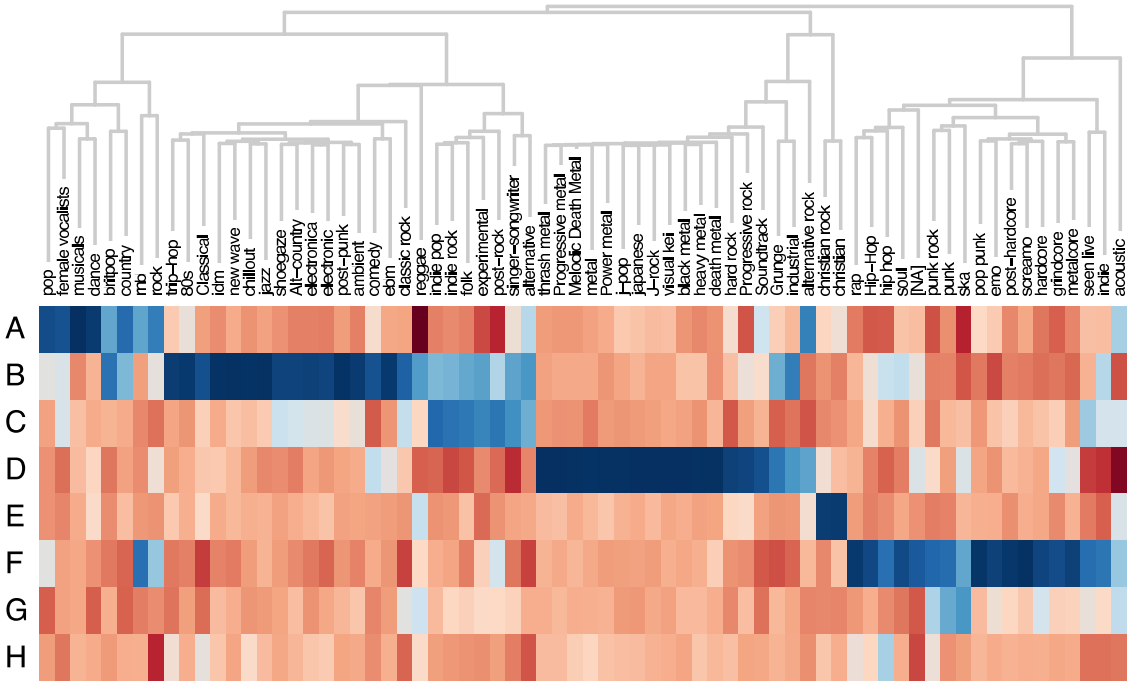


Figure 4.5: Last.fm is an Internet site that learns the musical taste of its members on the basis of examples, and then constructs a personalized, radio-like music feed. The web site also has a richer array of services, including a possibility to announce friendships with other users. The friendship network alone, when divided into components, reveals musical structures, because the music tastes of friends tend to be similar. Here the latent components found by our model were afterwards correlated with user's listening habits; songs are aggregated by tags given to them. Tags are intuitively grouped into genres. The network has 147,000 nodes and 353,000 links, but the running time with an efficient implementation by our collaborators at Xtract Ltd. was just 8.4 hours.

Bioinformatics and Neuroinformatics

Chapter 5

Bioinformatics

Samuel Kaski, Janne Nikkilä, Merja Oja, Jaakko Peltonen, Jarkko Venna, Antti Ajanki, Andrey Ermolov, Ilkka Huopaniemi, Arto Klami, Leo Lahti, Jarkko Salojärvi, Abhishek Tripathi

5.1 Introduction

New so-called high-throughput measurement techniques have made possible genome-wide studies of gene function. Gene expression, gene regulation, protein content, protein interaction, and metabolic profiles can be measured and combined with the genetic sequence. The methods are used routinely in modern biology and medicine, and now the current challenge is to extract meaningful findings from the noisy and incomplete data masses, collected into both community resource and private data banks. The data needs to be analyzed, mined, understood, and taken into account in further experiments, which makes data analysis an integral part of biomedical research. Successful genome-wide analyses would allow a completely novel systems-level view into a biological organism.

Combining the different kinds of data produces new systems-level hypotheses about gene function and regulation, and ultimately functioning of biological organisms. We develop probabilistic modeling and statistical data analysis methods to advance this field. Our main novel contributions stem from the cross-breeding of the methodological basic research, in particular on Modeling of Relevance, and collaboration with top groups in Biology and Medicine. We have had long-standing collaboration with Laboratory of Cytomolecular Genetics (Prof. S. Knuutila) and Neuroscience Center (Prof. E. Castrén), University of Helsinki, University of Uppsala (Prof. J. Blomberg), Turku Centre for Biology (Doc. T. Aittokallio), VTT (Prof. M. Oresic), and smaller-scale collaboration with several other groups. During 2007 we started new projects with EBI, UK (A. Brazma) and Finnish CoE in Plant Signal Research, University of Helsinki (Prof. J. Kangasjärvi) with promising results that will be reported in the next biennial report.

In 2006 we started a new conference series in collaboration with Prof. E. Ukkonen and J. Rousu of University of Helsinki. The conference "Probabilistic Modeling and Machine Learning in Structural and Systems Biology" inspired a special issue in a main journal, and yearly conferences in Evry, France, in 2007, and in Belgium in 2008.

References

[1] Juho Rousu, Samuel Kaski, and Esko Ukkonen, editors. *Probabilistic Modeling and Machine Learning in Structural and Systems Biology. Workshop Proceedings; Tuusula, Finland, June 17-18*. Helsinki, Finland, 2006.

[2] Samuel Kaski, Juho Rousu, and Esko Ukkonen. Probabilistic modeling and machine learning in structural and systems biology. *BMC Bioinformatics*, 8(Suppl 2):S1, 2007.

5.2 Translational medicine on metabolical level

Translational medicine is a research field which attempts to more directly bring basic research findings to clinical practice. One of the necessary steps of this process is to translate inferences made on the molecular level, for example about metabolites, in model organisms into inferences about humans. Such translation is extremely challenging and the existing knowledge, if there is any, is currently largely tacit and only known to experts of the specific disease and model organism.

Metabolomics is the study of the set of all metabolites found in a sample tissue. Metabolite concentrations are affected strongly by diseases and drugs, and hence they complement the genomic, proteomic, and transcriptomic measurements in an excellent way, in studies of the biological state of an organism.

We are in the process of developing new computational methods for translational medicine, for mapping between the observed metabolomics data from model organisms and humans. In project TRANSCENDO we apply the methods to studies of the emergence of Type I diabetes, by computing mappings between non-obese diabetic (NOD) mice and children, and between the effects of a disease in several tissues. The project is collaboration within a consortium involving computational systems biology (Matej Oresic, VTT), semantic modelling (Antti Pesonen, VTT), probabilistic modelling (us), and pharmacology and animal models of metabolic disease (Eriika Savontaus, University of Turku).

Metabolomic development in humans

Metabolic development of children and its differences between the genders is not yet well understood. These dynamic changes may, however, affect strongly the susceptibility to diseases and the responses to drugs.

We are studying a metabolomic data set derived from a collection of blood samples collected during the first years of life from boys and girls. We assume that the metabolic profiles are generated by a set of unobserved metabolic states, and we model those states and the data with a Hidden Markov Model (HMM). HMM fits the assumption of latent states very well and is easy to compute and interpret. Moreover, HMM provides a way for probabilistic re-alignment of the time series, which takes into account the individual variation in the dynamics. Simulations have indicated that HMMs can separate the boys' and girls' metabolic states more efficiently apart than traditional linear method; classification accuracy is 73% for HMM, and under 60% for linear methods. Figure 5.1 presents the model structures for girls and boys.

Disease-related dependencies between multiple tissues

A common setting in medical research is that a disease may be mainly located in a specific organ, for example in lungs, but it indirectly affects multiple tissues. Giving drugs to patients induces an analogous setup: the drugs may affect multiple other tissues in addition to the target tissue (and hence disease). We are developing new methods for discovering the disease-related metabolic dependencies between the multiple tissues, with the goal of revealing potential side effects of the diseases and drugs.

In practice, we have metabolomics data from mice belonging to 4 classes: healthy and untreated, sick and untreated, healthy and treated, sick and treated. A fast and straightforward way of digging out disease-related dependencies is to first find disease-related aspects with partial least-squares classifiers, and then dependencies with canonical correlation analyses and more straightforward correlations between contributing metabolites.

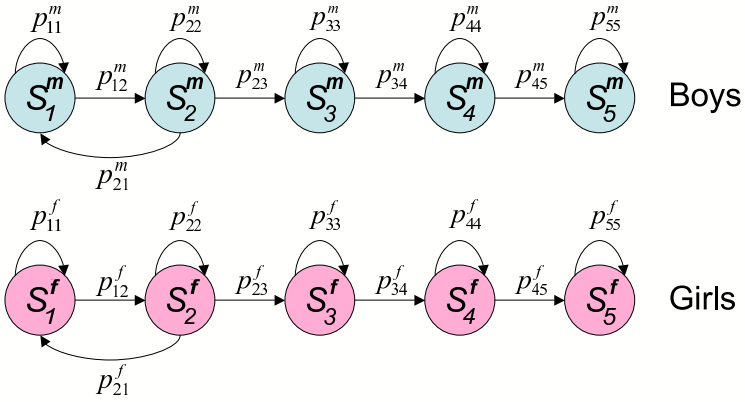


Figure 5.1: HMM models for metabolic states in boys and girls. The nodes represent hidden metabolic states, and the arrows possible transitions. Note that the states form a chain in order to force the models to focus on progressive changes in metabolite concentrations.

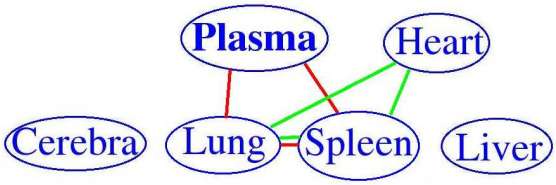


Figure 5.2: Disease-related dependencies between tissues before treatment (red), and after treatment (green). The disease is located in the lungs so the dependencies between lungs and plasma and spleen are logical, but note that after the treatment the dependency with plasma disappears and a dependency to heart emerges. This might be a sign of a side effect of the treatment.

This multivariate approach complements the traditional metabolite-wise linear models. Figure 5.2 shows the dependencies found between tissues before and after drug treatment.

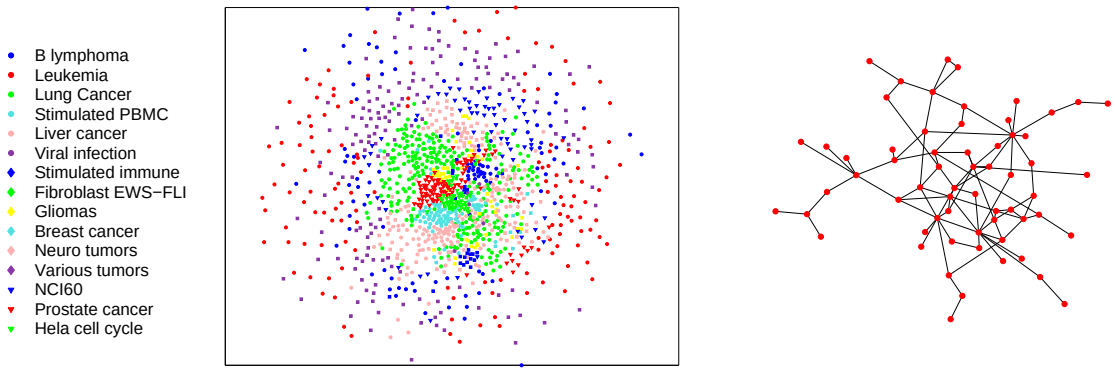


Figure 5.3: *Left:* Sample visualization of a gene expression atlas of cancer samples by curvilinear component analysis. Each dot denotes one microarray; the colors show the cancer class of the sample. *Right:* Part of yeast gene regulatory interaction network visualized by local multidimensional scaling.

5.3 Visualizing gene expression and interaction data

A large community-resource or private gene expression databank consists of numerous data sets submitted by several parties. They may have been measured for different purposes, with different treatments and methods in different laboratories. Several such databanks have been established and they continue to grow. A key challenge is how to best use the databanks to support further research. Currently information in these databanks is accessed using queries on the imperfect meta-data, that is, textual annotations and descriptions. In the future more sophisticated search methods, that take the actual data into account, are needed. Our study [2] aimed at comparing the different methods applicable as a visual interface that reveals similarities of data sets.

We compared several different visualization methods in the task of visualizing a large collection of gene expression arrays. Several new methods have been recently proposed for the estimation of data mani- folds or embeddings, but they have so far not been compared in the task of visualization. In visualizations the dimensionality is constrained, in addition to the data itself, by the presentation medium. It turned out that an older method, curvilinear components analysis, outperforms the new ones in terms of trustworthiness of the projections. Even though the standard preprocessing methods still need to be improved to make measurements of different labs and platforms more commensurable, the good news is that the visualized overview, expression atlas, reveals many of the cancer subsets (Fig. 5.3). Hence, we conclude that dimensionality reduction even from 1339 to 2 can produce a useful interface to gene expression databanks.

Biological high-throughput data sets can also be visualized as graphs that represent the relations between the biological entities. We applied our visualization methods for visualizing gene interaction graphs, and showed that Local Multidimensional Scaling performs very well in this task (Fig. 5.3; [1]).

References

[1] Jarkko Venna and Samuel Kaski. Visualizing Gene Interaction Graphs with Local Multidimensional Scaling In *Proceedings of ESANN'06, 14th European Symposium on Artificial Neural Networks*, pages 557–562, d-side, Evere, Belgium, 2006.

[2] Jarkko Venna and Samuel Kaski. Comparison of visualization methods for an atlas of gene expression data sets *Information Visualization*, 6:139–154, 2007.

5.4 Fusion of gene expression and other biological data sets

While analysis of gene expression data is a corner stone in modern bioinformatics, it is not a sufficient description of cellular state. The cell is an extremely complex system, and gene expression is only a partial view, among all the other omics. Only integration of information from multiple sources can reveal the true potential of the modern high-throughput measurement methods, such as gene expression data.

Integration is not trivial since the data types and scales can vary dramatically. Moreover, what is a proper way of doing the integration depends on the analysis task. Our main novel contribution has been to develop and apply new methods for searching for relevant features by combining data sources (described in Section Modeling of Relevance). We have additionally developed more specific methods for taking into account the known regulatory and context variables in modeling gene expression.

Relevant features through data fusion

We consider a data fusion problem of combining two or more data sources where each source consists of vector-valued measurements from the same object or entities but on different variables. The task is to include only those aspects which are *mutually* informative of each other. This task of including only shared aspects of data sources is motivated through two interrelated lines of thought. The first is noise reduction. If the data sources are measurements of the same entity corrupted by independent noise, discarding source-specific aspects will discard the noise and leave the shared properties that describe the shared entity. The second motivation is to analyze what is interesting in the data. One example is the study of activation profiles of yeast genes in several stressful treatments in the task of defining yeast stress response. In this example what is in common in the sources is what we are really interested in. The "noise" may be very structured; its definition is simply that it is source-specific.

A recent application is search for asbestos-related effects in gene expression by combining several cell lines [3].

We recently showed that there is a simple and computationally fast way of doing data fusion such that shared, relevant features are retained and source-specific noise is discarded [2]. The method is based on the classical canonical correlation analysis; it is surprising that there are still new practically important uses for so old methods! The method has been applied to several gene expression studies: classification of cell cycle regulated genes in yeast, identification of differentially expressed genes in leukemia, and defining stress response in yeast. The software package is available at `http://www.cis.hut.fi/projects/mi/software/drCCA/`.

Modeling context specific gene expression regulation

The biological state of the cell is to a large part defined by which genes are expressed at a certain moment or under certain environmental conditions. Regulation of gene expression is thus the key to understanding, for example, the reasons why some cells become transformed to cancer cells. Regulation of expression has been under intensive study during the past years, but analysis with statistical models has proved to be extremely difficult because the sample sizes are always small due to high measurement costs. The effective sample sizes become even smaller when analyzing context specific regulation, where the data becomes divided according to the context or experimental setup. We have introduced ways of context-specific modeling with one of the most often used model families, the Bays networks.

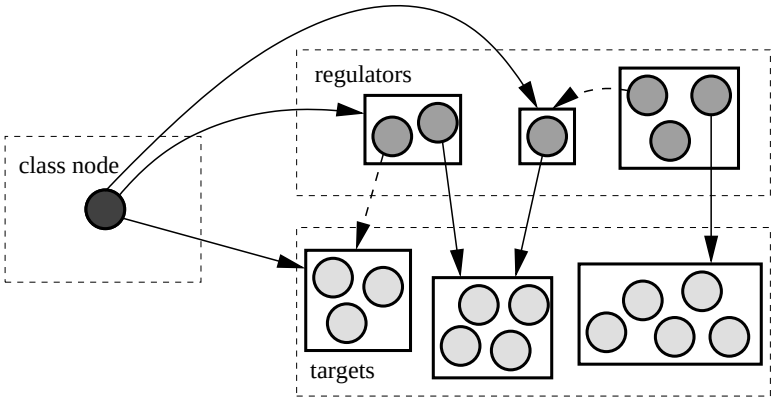


Figure 5.4: The structure of condition-dependent Bayesian network. Similarly behaving genes are grouped into modules. The edges depict the regulatory interactions. The dashed edges indicate interactions that are active in one of the conditions only. Genes linked from the class node may behave differently in different conditions.

The gene regulatory relationships form a complex network in which genes can be regulated by multiple regulators or through long chains of regulatory interactions. The regulatory network adapts to the conditions outside the cell by activating or stopping regulatory interactions in response to changes in the environment. We have studied regulatory networks in yeast with new *condition-dependent Bayesian network* [1]. The data has been divided into several conditions or contexts indicated by a context or class variable that is treated as a covariate. The model has the novel capability of identifying interactions that are active only in subset of conditions. The output of the method is a graphical representation of the network where the possible condition-dependent interactions are highlighted. Figure 5.4 depicts a conceptual example of a condition-dependent network.

We analyzed the regulation in yeast cultures which had been subjected to normal and stressful growth conditions. The method identified 25 regulators which are active only in stressful conditions. The majority of them (20 out of 25) have been annotated stress-related in the literature. The rest are new potential stress regulators.

References

[1] Antti Ajanki, Janne Nikkilä, and Samuel Kaski. Discovering condition-dependent Bayesian networks for gene regulation. In *Proceedings of Fifth IEEE International Workshop on Genomic Signal Processing and Statistics*, 2007.

[2] Abhishek Tripathi, Arto Klami, and Samuel Kaski. Simple integrative preprocessing preserves what is shared in data sources. *BMC Bioinformatics*, 9:111, 2008.

[3] Penny Nymark, Pamela M Lindholm, Mikko V Korpela, Leo Lahti, Salla Ruosaari, Samuel Kaski, Jaakko Hollmen, Sisko Anttila, Vuokko L Kinnula, and Sakari Knuutila. Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines. *BMC Genomics*, 8:62, 2007.

5.5 Human endogenous retroviruses

The human genome includes surviving traces of ancient infections by retroviruses that have become fixed to human DNA. These surviving traces are called *human endogenous retroviruses* (HERVs). HERVs are interesting because they can express viral genes in human tissues, and because their presence in the genome may affect the functioning of nearby human genes. If ancient highly mutated elements are included, HERV sequences form 8% of the human genome [1].

In earlier research we had used Self-Organizing Maps to analyze the classification of HERVs into families [2]. In recent research we have moved to estimating the relative activities (expression levels) of the HERVs across several human tissues. We analyze activity for individual HERV sequences (rather than groups of sequences); this is vital for analyzing their individual control mechanisms and their possible roles in diseased and normal cell functions.

To find evidence of HERV activity, we use probabilistic modeling methods for expressed sequence tags (ESTs) gathered from public databases. We introduced a generative mixture model for EST sequences where each component of the mixture was associated with a particular HERV (see the top subfigure of Fig. 5.5). In our experiments we compared this rigorous model with a fast heuristic method; it turned out that the fast method performed reasonably accurately on simulated data, which made it possible to analyze very large HERV collections.

We first used the models to analyze overall activities across different tissues and conditions. In addition to comparisons on simulated data, we performed several experiments on real HERV data; the probabilistic method for a smaller and the fast method for a larger set having 2450 HERVs [3]. Lastly the probabilistic model was used to estimate tissue-specific expression of HERVs from the HML2 family [4].

Overall, 7% of the HERVs were estimated to be active; the majority of the HERV activities were previously unknown. HERVs with the retroviral *env* gene were found to be more often active than HERVs without *env*. We were also able to analyze which parts of the HERV sequences the EST data match to; see [4] and its supplementary material for figures. For the HERV family HML2, activity profiles of HERVs over tissues are shown in the bottom subfigure of Fig. 5.5; some of the HML2 HERVs display tissue-specific expression (e.g. activity in male reproductive tissues or in the brain).

References

[1] Eric S. Lander *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[2] Merja Oja, Göran O. Sperber, Jonas Blomberg, and Samuel Kaski. Self-organizing map-based discovery and visualization of human endogenous retroviral sequence groups. *International Journal of Neural Systems*, 15(3):163–179, 2005.

[3] Merja Oja, Jaakko Peltonen, Jonas Blomberg, and Samuel Kaski. Methods for estimating human endogenous retrovirus activities from EST databases. *BMC Bioinformatics*, 8(Suppl 2):S11, 2007.

[4] Merja Oja. In silico expression profiles of human endogeneous retroviruses. In *Proceedings of the Workshop on Pattern Recognition in Bioinformatics (PRIB 2007)*, 2007.

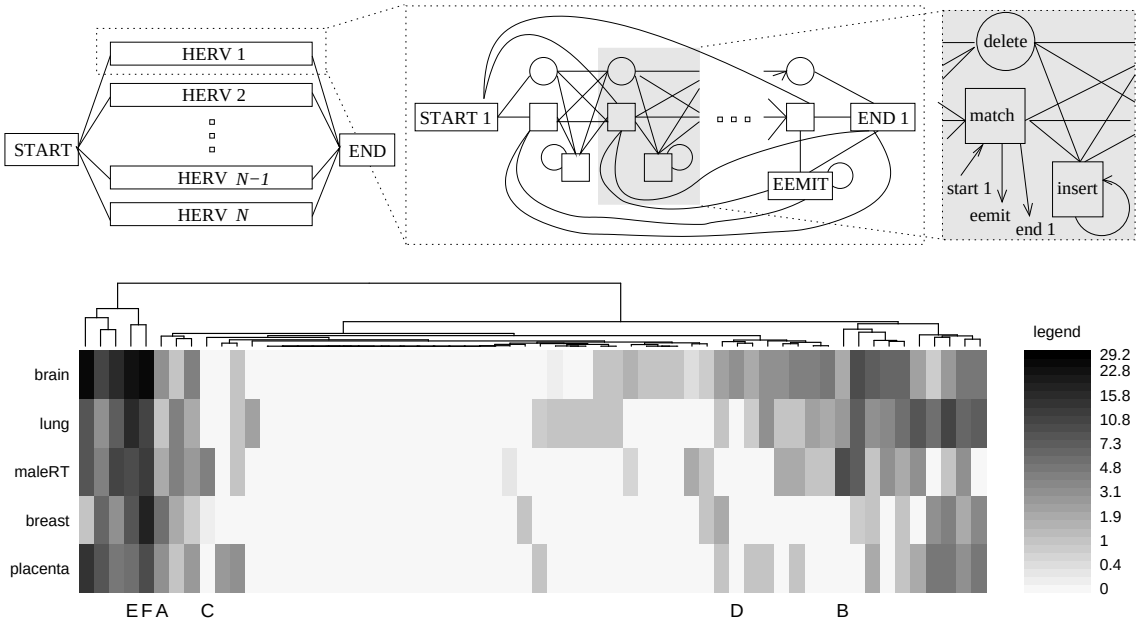


Figure 5.5: Top: the new probabilistic mixture model introduced for estimating activity of human endogeneous retroviruses (HERVs) from expressed sequence tags (ESTs). Bottom: activities of HERVs from the HML2 family in different tissues. Each column depicts the expression profile of an individual HERV sequence; the columns have been ordered by hierarchical clustering based on the profiles. Numbers next to the legend are probabilistic EST counts. Letters A-F at the bottom identify individual HERVs that have been analyzed in [4].

Chapter 6

Neuroinformatics

Ricardo Vigário, Jaakko Särelä, Sergey Borisov, Astrid Pietilä, Jan-Hendrik Schleimer, Jarkko Ylipaavalniemi, Alexander Ilin, Samuel Kaski, Eerika Savia, Erkki Oja

6.1 Introduction

Neuroinformatics has been defined as *the combination of neuroscience and information sciences to develop and apply advanced tools and approaches essential for a major advancement in understanding the structure and function of the brain.* Aside from the development of new tools, often the fields of application include the analysis and modelling of neuronal behaviour, as well as the efficient handling and mining of scientific databases. With the current configuration, the group aims at proposing algorithmic and methodological solutions for the analysis of elements and networks of functional brain activity, studying several kinds of communication mechanisms. These are to be applied in the understanding of ongoing brain activity, as well as responses to natural stimulation.

From a methodological viewpoint, the neuroinformatics group has been involved in studying certain properties of ICA, such as its reliability and applicability to the analysis of electrophysiological brain data (namely electroencephalograms, EEGs and magnetoencephalograms, MEG), as well as to functional magnetic resonance images (fMRI). Within the study of ICA reliability, subspace effects have been made evident, and their potential in functional network interpretability suggested (see Sec. 6.2).

Two explorative studies into functional brain networks have been started. One made explicit use of complex stimulation in fMRI, resulting in the detection of several networks of functional activity with clear interpretability (see Sec. 6.3). Another targeted phase synchrony, which is expected to play a central role in the communication within the central nervous system, as well as between this and the peripheral nervous system (see Sec. 6.4).

Several other topics have been researched in the field of biomedical signal processing, which are not thoroughly reported here. In particular, the denoising source separation framework (DSS) introduced earlier in the laboratory of computer and information science, has been used in the study of phonocardiographic signals, as well as in the investigation of different possible origins for high- and low-amplitude alpha-activity in EEG. We have as well studied measurement fMRI artefacts using a reliable ICA approach with a standard spherical phantom. All of these topics are collected in Sec. 6.5. The application of our methods to tissue segmentation in magnetic resonance imaging (MRI), to the detection of brain lesions will appear in the report of next biennial period.

Research reported in this section has been carried out in collaboration with experts in neuroscience and cardiology.

References

[1] Schleimer, J.-H., and R. Vigário. Clustering limit cycle oscillators by spectral analysis of the synchronisation matrix with an additional phase sensitive rotation. In *Proc. 17th Int. Conf. on Artificial Neural Networks (ICANN'2007)*, Porto, Portugal, pp. 944–953, 2007.

[2] Schleimer, J.-H., and R. Vigário. Order in Complex Systems of Nonlinear Oscillators: Phase Locked Subspaces. In *Proc. of 15th European Symposium on Artificial Neural Networks (ESANN'07)*, Bruges, Belgium, pp. 13–18, 2007.

[3] Ylipaavalniemi, J., E. Savia, R. Vigário, and S. Kaski. Functional elements and networks in fMRI. In *Proc. of 15th European Symposium on Artificial Neural Networks (ESANN'07)*, Bruges, Belgium, pp. 561–566, 2007.

[4] Ylipaavalniemi, J., and R. Vigário. Subspaces of Spatially Varying Independent Components in fMRI. In *Proc. 7th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA'2007)*, London, England, pp. 665–672, 2007.

[5] Schleimer, J.-H., and R. Vigário. Reference-based extraction of phase synchronous components. In *Proc. 16th Int. Conf. on Artificial Neural Networks (ICANN'2006)*, Athens, Greece, pp. 230–238, 2006.

[6] Borisov, S., A. Ilin, R. Vigário, and E. Oja. Comparison of BSS methods for the detection of α-activity components in EEG. In *Proc. 6th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA'2006)*, Charleston, South Carolina, USA, pp. 430–437, 2006.

[7] Pesonen, M., M. Laine, R. Vigário, and C. Krause. Brain oscillatory EEG responses reflect auditory memory functions. *abstr. 13th World Congress of Psychophysiology, International Organization of Psychophysiology (IOP'2006)*, Istanbul, Turkey, 2006.

[8] Pietilä, A., M. El-Segaier, R. Vigário, and E. Pesonen. Blind Source Separation of Cardiac Murmurs from Heart Recordings. In *Proc. 6th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA'2006)*, Charleston, South Carolina, USA, pp. 470–477, 2006.

[9] Ylipaavalniemi, J., S. Mattila, A. Tarkiainen, and R. Vigário. Brains and Phantoms: An ICA Study of fMRI. In *Proc. 6th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA'2006)*, Charleston, South Carolina, USA, pp. 503–510, 2006.

6.2 Reliable ICA and subspaces

In contrast to traditional hypothesis-driven methods, independent component analysis (ICA) is commonly used in functional magnetic resonance imaging (fMRI) studies to identify, in a blind manner, spatially independent elements of functional brain activity. Particularly, in studies using multi-modal stimuli or natural environments, where the brain responses are poorly predictable, and their individual elements may not be directly relatable to the given stimuli.

In earlier reported work, we have analysed the consistency of ICA estimates, by focusing on the spatial variability of the components. The optimization landscape of ICA is defined by structure of the data, noise, as well as the objective function used. The landscape can form elongated or branched valleys, containing many strong points, instead of singular local optima. Multiple runs of the ICA algorithm with varying random initial conditions and re-sampling allows to characterise the optimisation landscape and the robustness of the estimates.

Previous studies have analyzed the consistency of independent components, and suggested that some components can have a characteristic variability. The goal was to provide additional insight into the components, that is not possible to attain with single run approaches. Complex valleys can also be considered as separate subspaces, where statistical independence is not necessarily the best objective for decomposition.

We have now proposed a novel method for reliably identifying subspaces of functionally related independent components. We also proposed two approaches to further refine the decomposition into functionally meaningful components. One refinement method uses clustering, to distinguish the internal structure of the subspace. Another method is based on finding the coordinate system inside the subspace that maximally correlates with the temporal dynamics of the stimulation. The directions are found with canonical correlation analysis (CCA).

A study of subspaces was conducted on multi-modal fMRI recordings, including several forms of auditory stimulation. In the following figure, we can see a set of components, strongly related to auditory stimulation. Each component is consistent, appearing in all or most of the 100 runs. The mixing variability is also minimal. However, the spatial variance reveals a coincident location of variability, shared by all components. The variability links the components into a three dimensional subspace, even though ICA has consistently identified directions within the subspace.

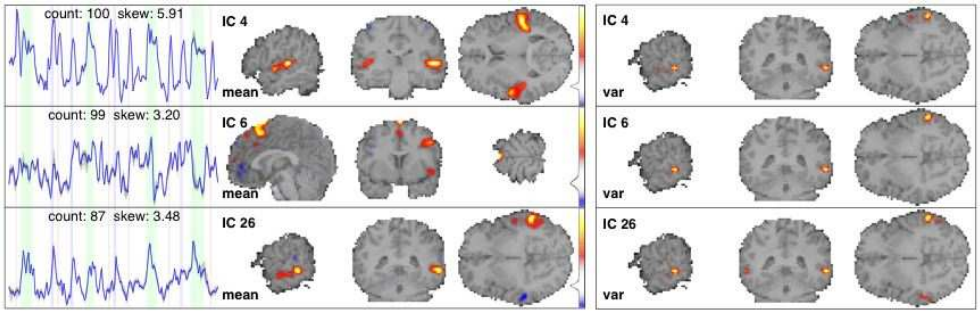


Figure 6.1: Tested approaches for alpha extraction from simulated data.

Within the study, we postulate that, based on spatial variance, components can be roughly divided into 3 classes: individual and consistent components, with distributed

variance due to noise; consistent members of a subspace, with focal variance coincident with the variance of the other members; and inconsistent subspaces, with variances coincident with their own mean. Such subspaces can provide information on networks of related activity in a purely data-driven manner. Criteria to disambiguate each subspace will then be of crucial relevance.

6.3 Towards brain correlates of natural stimuli

Natural stimuli are increasingly used in functional magnetic resonance imaging (fMRI) studies to imitate real-life situations. Consequently, challenges are created for novel analysis methods, including new machine learning tools. With natural stimuli it is no longer feasible to assume single features of the experimental design alone to account for the brain activity. Instead, relevant combinations of rich-enough stimulus features could explain the more complex activation patterns.

We proposed a novel two-step approach, where independent component analysis is first used to identify spatially independent brain processes, which we refer to as functional patterns. As the second step, temporal dependencies between stimuli and functional patterns are detected using dependency exploration methods. Our proposed framework looks for combinations of stimulus features and the corresponding combinations of functional patterns.

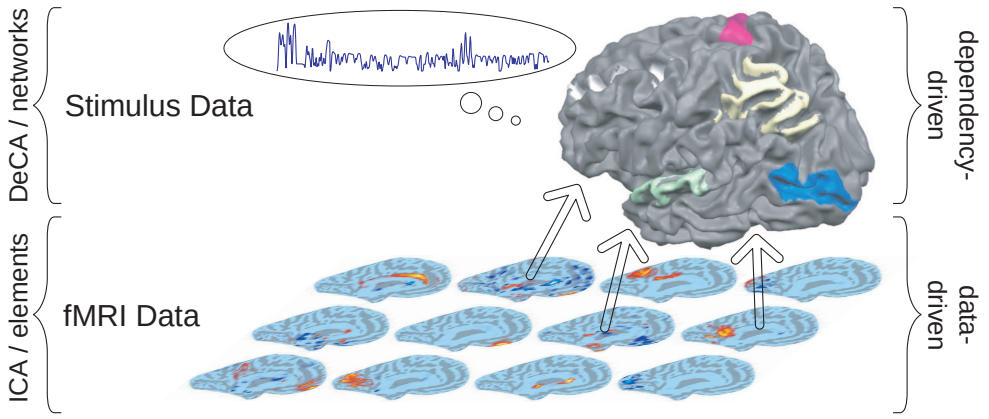
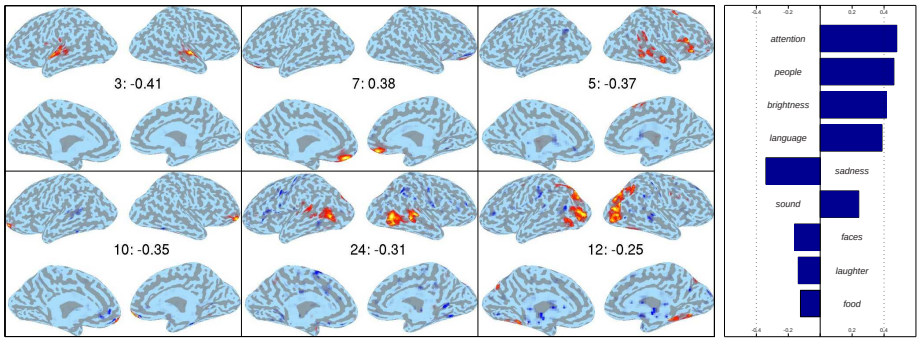


Figure 6.2: The proposed framework: elements of functional brain activity emerge from the data via ICA. Functional networks are revealed by DeCA, based on covariation between the elements and task goals, encoded as features.

This two-step approach was tested on fMRI recordings of brain responses to natural stimuli, consisting of a movie with 20 minutes duration. Rather subjective features were extracted from the movie, including labels such as *"attention"*, *"sadness"*, *"people"* or *"laughter"*.

As an illustrative example, we can look into a network comprising brain areas that individually correspond to, *e.g.*, auditory (IC3), visual (IC12), and multi-modal integration (IC24). This suggests that the functional role of the whole network is related to combining information from many sensory inputs. Indeed, the four highest scoring features of the dependent component are *attention*, *people*, *brightness* and *language*.

The found networks seem plausible, considering the limited and very subjective nature of the available stimulus features. Some elements were a part of several networks, with different functional contribution to each networks common task. A more controlled study has been carried out since, to verify the results and to further develop the approach. These will be reported in the next biennial report.



(a) (b)

Figure 6.3: (a) The 6 ICs corresponding to the highest loadings in the first dependent component. The loading value is shown in the middle of each square. (b) Respective loadings of the stimulus features.

6.4 Synchrony exploration

Interest in phase synchronisation phenomena has a long history, when studying the interaction of complex, natural or artificial, dynamic systems. Although not completely adopted, synchronisation was attributed a role in the interplay between different parts of the central nervous system as well as across central and peripheral nervous systems. Such phenomena can be quantified by the phase locking factor (PLF), which requires knowledge of the instantaneous phase of an observed signal.

Linear sources separation methods treat scenarios in which measurements do not represent direct observations of the dynamics, but rather superpositions of underlying latent processes. Such a mixing process can cause spuriously high PLF's between the measurements, and camouflage the phase locking to a provided reference signal. Essentially, synchronisation is either caused by a common input or by interactions between neurons.

Reference-based approach

The PLF between a linear projection of the data and a reference can be maximised as an optimisation criterion, revealing the most synchronous source component present in the data, with its corresponding amplitude. This is possible despite the amplitude distributions being Gaussian, or the signals being statistically dependent, common assumptions in blind sources separation techniques without a-priori knowledge, e.g. in form of a reference signal.

We first addressed this reference-based problem, and proposed a new algorithm capable of extracting sources phase-locked with a reference. In the following illustration one can see the efficiency of such a method. The sources, depicted on the right frame, were chosen so that neither high-order statistics methods, e.g., FastICA, nor methods based on temporal decorrelation, e.g., SOBI would perform the desired source estimation.

We tested this approach on MEG recordings, with a 306-sensor Vectorview neuromagnetometer, together with left and right hand EMG's. The subject was instructed to simultaneously keep isometric contraction in left and right hand muscles, using a special squeezing device. We then used the right hand EMG as a reference for the phase exploration into the MEG recordings. The results achieved agreed with early studies performed in the same recordings.

We also addressed the "internal neuronal synchronisation" problem, where no clear reference is available, proposing to cluster a population of oscillators into segregated sub-

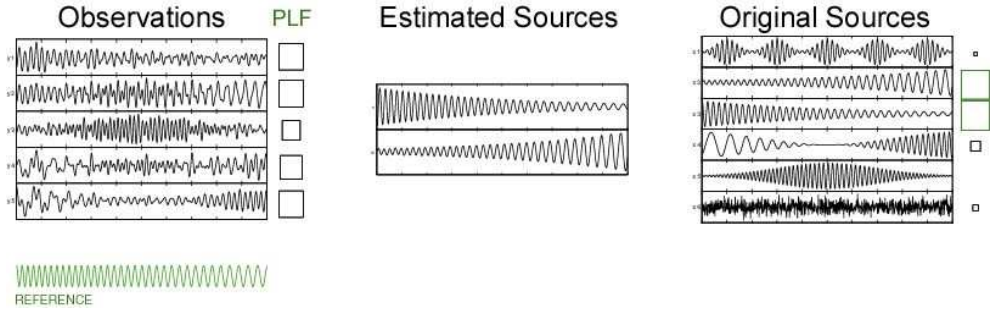
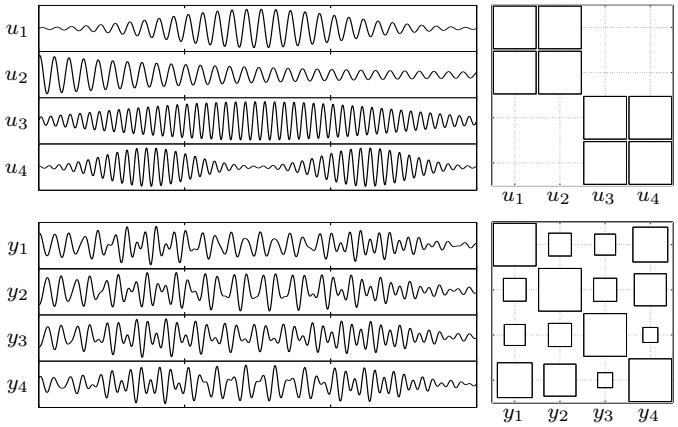


Figure 6.4: Six signals, of which only two are locked in phase.



(a) Oscillators with distinct amplitudes, grouped in synchronous subspaces. PLFs are depicted as a Hinton diagram. ($P = 0.08$)

(b) Mixtures with an overall spuriously phase locking. ($P = 3.66$)

populations, exhibiting high internal interactions. Approaches to solve this problem have often assumed different frequencies for the various sub-populations, usually neglecting phase information. These assumptions pose a restriction to the analysis of the dynamic world of natural systems, where communication can be unrelated to the natural frequency of the constituent oscillators.

Our solution makes explicit use of phase information, extracted from known models of physical interactions. The approach relies on a post-rotation of the eigenvectors of the synchronisation matrix.

With simulations, we show the effect of the post-rotation, in the estimation of underlying sources for which their frequency has been drawn from a global distribution. In neurobiological terms, this means that the neuron's system parameters, which determine its natural frequency, do not depend on the synaptic connections it has formed. In such formulation, frequency can not be used to identify the sources anymore. Phase clustering is then crucial for the task.

We have also proposed a method to reveal phase-locked subspaces, based on a concept of order in complex systems of nonlinear oscillators. Any order parameter quantifying the degree of organisation in a physical system can be studied in connection to source extraction algorithms. Independent component analysis, by minimising the mutual information of the sources, falls into that line of thought, since it can be interpreted as searching components with low complexity. Complexity pursuit, a modification minimising Kolmogorov complexity, is a further example.

Using such concept of order, we designed an algorithm capable of revealing subspaces of oscillators such that: oscillators of the same subspace are completely phase locked; whereas between subspaces there is no Phase locking. The following illustrations exemplify the algorithm's performance. Estimated sources coincide with the true ones.

6.5 Overview of other topics

Within the duration of this biennial report, several research topics have been addressed with a more prospective view. Some will be the subject of more thorough development in further reports, whereas other will stay as simple case studies. This section reviews four of those.

Extraction of alpha activity

Following earlier work on the characterisation of low- and high-amplitude alpha brain activity, we tested a two-step blind source separation approach for the extraction of such rhythms from ongoing EEG. The method comprised a denoising stage, performed by DSS, followed by either a high-order statistical independent component analysis source estimation, FastICA, or a temporal decorrelation one, TDSEP.

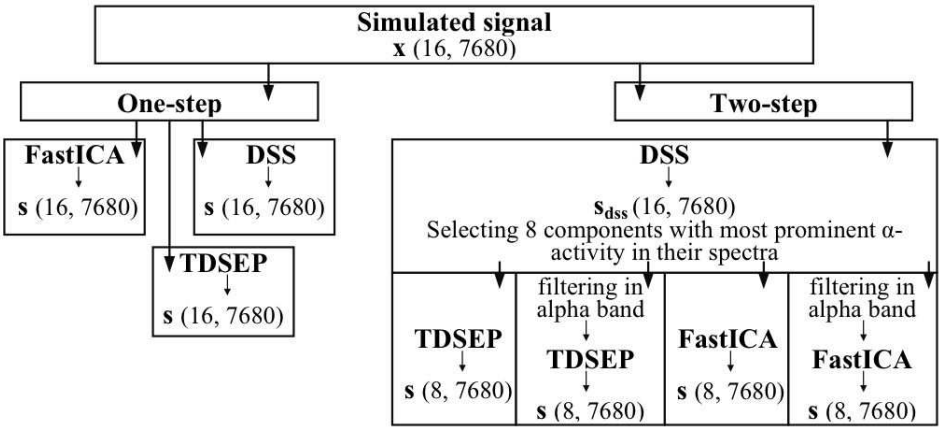


Figure 6.5: Tested approaches for alpha extraction from simulated data.

The main findings are that denoising has, as expected, a positive effect in rendering the subsequent source separation algorithms more efficient. In addition, we observed that high-order statistics ICA was more adequate in such separation than TDSEP, in spite of the latter being particularly suited for dealing with temporally structured sources. A targeting α-filter, placed between the denoising and the TDSEP modules, resulted in good estimates, rendering the combination rather efficient. Such filtering seems to not affect significantly FastICA.

Artefact removal in ERD/ERS study

Still within the rhythmic activity of the brain, we participated in a study of brain oscillatory EEG responses to auditory memory functions. The analysis concentrated on event related de-synchronisation and synchronisation (ERD and ERS, respectively), in the theta and alpha frequency ranges for ERS and also in beta for ERD.

The outcomes of that study suggested that theta frequency ERS responses may be associated with working memory functions, whereas alpha ERD/ERS responses robustly dissociate between auditory memory encoding and recognition.

ICA showed to be crucial in denoising the raw ongoing EEG, prior to wavelet processing. Several subjects displayed considerable artefacts that rendered most of the event-related responses virtually unusable.

BSS of cardiac murmurs

A significant percentage of young children present cardiac murmurs. However, only one percent of them are caused by a congenital heart defect; others are physiological. An automated system for an initial recording and analysis of the cardiac sounds could enable the primary care physicians to make the initial diagnosis and thus decrease the workload of the specialised health care system. independent component analysis source estimation, FastICA, or a temporal decorrelation one, TDSEP.

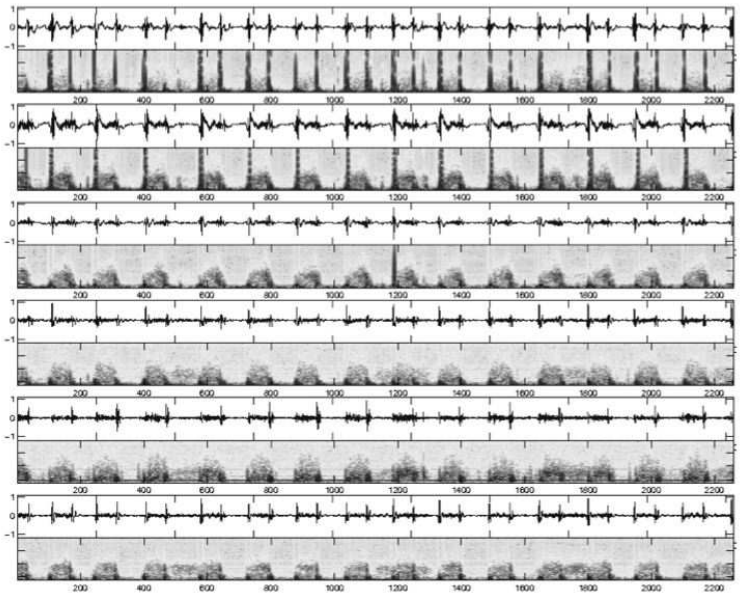


Figure 6.6: Six-channel PCG recordings from a patient, together with their spectrograms. The S1 and S2 are clearly visible in the first spectrogram as periodic pairs of vertical bars covering all the frequencies. Murmurs are visible in the systole, between the S1 and S2, present on all six recordings.

The first step to such analysis is the identification of the different components of the cardiac cycle, with particular emphasis to the separation of the murmurs. We have proposed a new methodological framework to address this issue, combining ICA and DSS. independent component analysis source estimation, FastICA, or a temporal decorrelation one, TDSEP.

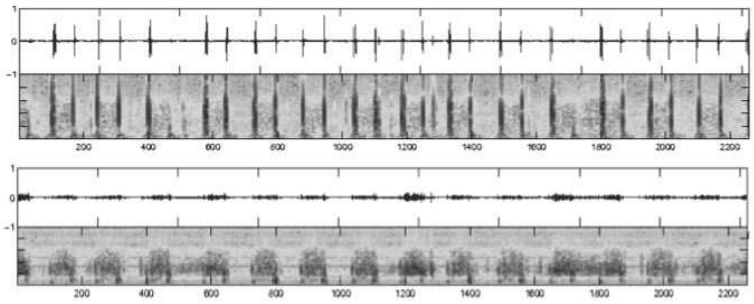


Figure 6.7: Heart sounds S1 and S2, clearly isolated from all other signals. In the second frame are uncontaminated murmurs.

Using such approach, we have been able to isolate rather efficiently the murmurs, as well as heart sounds S1 and S2 and artefacts such as voices recorded during the measurements.

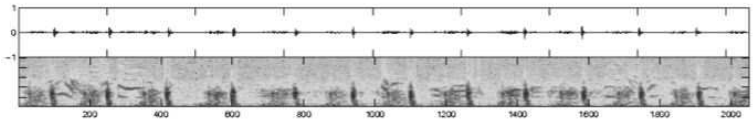


Figure 6.8: Speech artefacts present in PCG recordings. Formant structures are clearly visible.

With the aforementioned results, the collaboration with the Lund University Hospital, Sweden, has been strengthened, and further research outcomes are expected in the next reports.

Phantom study in fMRI

Phantom measurements are routinely used for verifying and calibrating the quality of MRI machinery. However, data-driven analysis of phantom fMRI data has been largely overlooked, possibly due to the lack of a method for assessing the reliability of the solutions. We have now used a reliable ICA approach to such analysis, and revealed evidence for possible misinterpretations in ICA studies with real subjects.

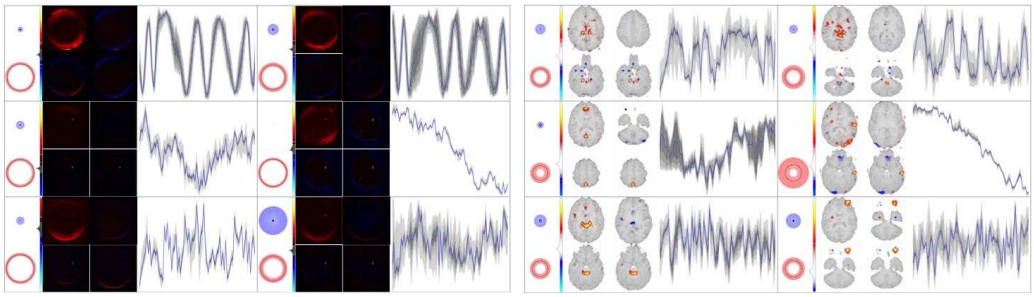


Figure 6.9: Reliable independent components, extracted from fMRI of a spherical phantom (a), and a real subject (b). Corresponding temporal 'activation' patterns are shown on the right of each estimate.

Several independent components found on a real subject presented a temporal structure that follows clearly that of phantoms. We speculate that methods other than ICA can also suffer from a similar kind of misinterpretation. We therefore suggest the need for a better understanding of the artificial, scanner- or environmentally-induced artefacts, prior to the automatic analysis of any fMRI recording. A comparison between real brain ICA and phantom-based decompositions may help in the validation of the estimated components.

Multimodal interfaces

Chapter 7

Content-based information retrieval and analysis

Erkki Oja, Jorma Laaksonen, Markus Koskela, Ville Viitaniemi, Zhirong Yang, Mats Sjöberg, Hannes Muurinen

7.1 Introduction

Content-based image or information retrieval (CBIR) has been a subject of intensive research effort for more than a decade now. Content-based retrieval of images differs from many of its neighboring research disciplines in computer vision due to one notable fact: human subjectivity cannot totally be isolated from the use and evaluation of CBIR systems.

In our PicSOM[1] CBIR system, parallel Self-Organizing Maps (SOMs) have been trained with separate data sets obtained from the multimodal object data with different feature extraction techniques. The different SOMs and their underlying feature extraction schemes impose different similarity functions on the images, videos, texts and other media objects. In the PicSOM approach, the system is able to discover those of the parallel SOMs that provide the most valuable information for retrieving relevant objects in each particular query.

7.2 Benchmark tasks of natural image content analysis

In the course of previous years we have outlined and implemented our generic PicSOM system architecture for image and information retrieval tasks. The architecture is based on extraction of numerous different features from the feature descriptors from the information objects, performing inference separately based on each feature, and fusing the partial inferences. The architecture supports hierarchical organization of the information objects. In the case of image analysis, the hierarchy is used to describe the decomposition of images into segments.

We have investigated how our architecture can be applied to various benchmark tasks concerning generic domain photographic images. While individual components of the architecture have been improved during the studies, the general architecture has proven to be successful. The improvements include the incorporation of new feature extraction methods, most notably the Scale-Invariant Feature Transform (SIFT) features calculated from interest points, the use of Support Vector Machines (SVMs) as an alternative to SOMs as the classification method, and alternative early and late feature feature fusion methods.

Our group has participated in the annual PASCAL FP6 NoE Visual Object Classes (VOC) Challenges [1, 2]. The material of the Challenges consists of photographic images of natural scenes containing objects from predefined object classes. In 2006 there were approximately 5000 images and ten object classes, including objects such as “bicycle”,”bus”,”cat” and “cow”. For the 2007 Challenge the number of images and object classes were both doubled. The Challenge included the classification task, ie. the determination whether an object of a particular class appears in the image, and the detection task for the object’s bounding box. In addition, the 2007 Challenge also included a novel competition of pixel-wise object segmentation. Our performance in the Challenge has been satisfactory, the highlights being the best segmentation accuracy and the fourth best classification performance in the 2007 Challenge.

For the VOC benchmarks we have investigated and analyzed techniques of automatic image segmentation, especially in [3]. The devised techniques have been fundamental for performing the bounding box detection tasks. However, for the classification task the usefulness of segmented images does not currently seem to be competitive against state-of-the-art global image analysis techniques. Partly this is due to the strong correlation of

[1] http://www.cis.hut.fi/picsom

Figure 7.1: Images of the VOC2006 image collection shown together with those patches that on the collection level contribute most to the classification of the images as a "bus", "cow" and "motorbike".

actual target objects and the background both in the challenge databases and in natural images in general, which diminishes the advantage from focusing analysis exclusively to specific image locations. This effect is illustrated in Figure 7.1 where we have highlighted the image patches that contribute most to the decision of the image containing a particular object in the classification task.

Other benchmark tasks we have studied include the ImageCLEF 2006 object annotation task, which we analyzed outside the competition, and the ImageCLEF 2007 object retrieval task, in which our results were clearly the best of the campaign submissions. We have also applied our CBIR system to benchmark tasks of automatic image annotation, performing clearly better than numerous state-of-the-art methods reported in literature [4, 5].

7.3 Interactive facial image retrieval

It is often desired to search for an image depicting a person only through an eyewitness' recalling about the appearance. Interactive computer-based systems for this purpose, however, confront the problem of evaluation fatigue due to time-consuming retrieval. We have addressed this problem by extending our PicSOM CBIR system to emphasize the early occurrence of the first subject image. Partial relevance criteria provide a common language understood by both the human user and the computer system. In addition to filtering by ground truth and hard classifier predictions, we have proposed Discriminative Self-Organizing Maps (DSOMs) [6] to adaptively learn the partial relevances.

A straightforward method to obtain DSOMs is to employ discriminant analysis as a preprocessing step before normal SOM training. We have applied the widely used method, PCA+LDA, in pattern recognition as our baseline. Furthermore, we have adapted the Informative Discriminant Analysis (IDA) to maximize the discrimination for more complicated distributions. Our Parzen Discriminant Analysis [7] regularizes the IDA objective by emphasizing the prior of piecewise smoothness in images. Both LDA and our PDA have been extended for handling fuzzy cases. The original IDA optimization algorithm is computationally expensive. We have presented three acceleration strategies [8]: First, the computation cost of batch gradients is reduced by using matrix multiplication. Second, the updates follow an geodesic flow in the Stiefel manifold without Givens reparameterization. Third, a more efficient leading direction is calculated by preserving only the principal whitened components of the batch gradient at each iteration.

Simulations have been performed on the FERET database. We have provided a query example (Figure 7.2) and also presented a quantitative study on the advantage in terms of the first subject hit and retrieval precisions at various recall levels [6].

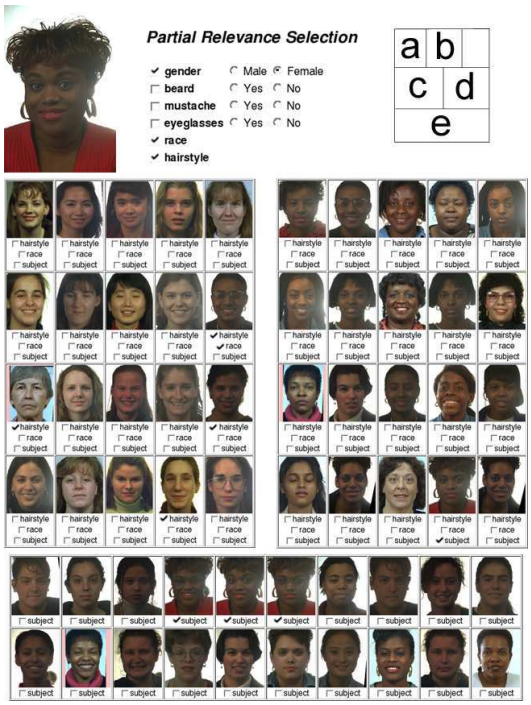


Figure 7.2: A query example using the PicSOM system: (a) the target person; (b) specifying partial relevance; the displayed images in the first phase, with (c) the first round and (d) the second round; (e) the images displayed in the first round of the second phase.

7.4 Content analysis and change detection in earth observation images

Earth observation (EO) data volumes are growing rapidly, with an increase in both the number of satellite sensors and in their resolutions. Yet, it is estimated that only 5% of all EO data collected up to now has been used. Therefore, traditional remote sensing archiving systems – with queries made typically on sensor type, geographical extents or acquisition date – could become obsolete as the amount of data to be stored, accessed and processed explodes. Using image content indexing would allow a more efficient use of these databases. This has led to the emergence of content-based image retrieval systems for archive management of remote sensing images and for annotation or interpretation of satellite images. In co-operation with the VTT Technical Research Centre of Finland, we have applied the PicSOM system for analysis of multispectral and polarimetric radar (PolSAR) satellite images divided in small patches or *imagelets*.

With the high-resolution optical images the aim has been to detect man-made structures and changes on the studied land cover. Fusion of panchromatic and multispectral information was done conveniently within the PicSOM framework, in which several SOMs are trained in parallel, one SOM per feature. Qualitative and quantitative evaluation of the methods were carried out for man-made structure detection and change detection, using partially labeled datasets. The results were encouraging, considering that a totally new approach was presented to the challenging problem of change detection in very high-resolution images [9]. Possible applications of this work are high-resolution satellite image annotation and monitoring of sensitive areas for undeclared human activity, both in an interactive way.

With the radar images, the availability of dual-polarization and fully-polarimetric data, instead of earlier single-polarization data, will in the near future enable a deeper analysis of backscattering processes. This development will in turn pave the way for many new applications for spaceborne SAR data. At the same time, these satellite missions generate a huge amount of data at a higher resolution than previous spaceborne SAR sensors. It is still quite unclear what low-level features will be the most efficient ones for the automatic content analysis of the satellite polarimetric SAR data. In our research [10] we have compared six different types of polarimetric features and their different postprocessings, including averages and histograms, to gain quantitative knowledge of their suitability for the land cover classification and change detection tasks. The results proved that different features are most discriminative for different land cover types, and the best overall performance can be obtained by using a proper combination of them.

Figure 7.3: 100×100-pixel optical and 16×16-pixel SAR (Pauli decomposition) imagelets.

7.5 Multimodal hierarchical objects in video retrieval

The basic ideas of content-based retrieval of visual data can be expanded to multimodal data, where we consider multimodal objects, for example video or images with textual metadata. The PicSOM system has been extended to support general multimodal hierarchical objects and to provide a method for relevance sharing between these objects [11]. For example a web page with text, embedded images and links to other web pages can be modeled as a hierarchical object tree with the web page as the parent object and the text, links and images as children objects. The relevance assessments originally received from user feedback will then be transferred from the object to its parents, children and siblings. For example, if we want to search for an image of a cat from a multimedia message database, we can let the system compare not only the images, but also the related textual objects. If the reference message text contains the word "cat" we can find images which are not necessarily visually similar, but have related texts containing the same keyword.

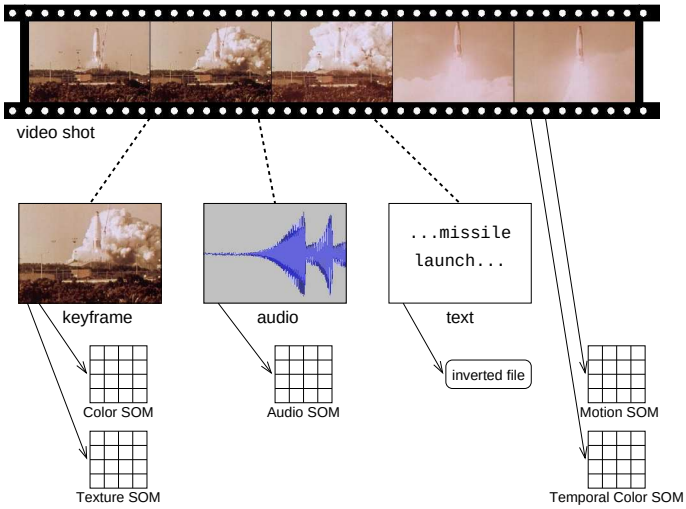


Figure 7.4: The hierarchy of video and multimodal SOMs.

The multimodal hierarchy used for indexing video shots and supporting multimodal fusion between the different modalities is illustrated in Fig. 7.4. The video shot itself is considered as the main or parent object in the tree structure. The keyframes (one or more) associated with the shot, the audio track, and text obtained with automatic speech recognition are linked as children of the parent object. All object modalities may have one or more SOMs or other feature indices, and thus all objects in the hierarchy can have links to a set of associated feature indices.

A common approach to semantic video retrieval is to combine separate retrieval results obtained with low-level visual features and text-based search. The relative weights of these sub-results are specified based on e.g. validation queries or query categorization.

An important catalyst for research in video retrieval is provided by the annual TREC Video Retrieval Evaluation (TRECVID) workshop. The goal of the workshop series is to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested to compare their results. The search task in TRECVID models the task of an intelligence analyst who is looking for specific segments of video containing persons, objects, events, locations, etc. of current interest. The task is defined as follows: given a search test collection and a multimedia statement of information need, return a ranked list of shots which best satisfy the need. We have successfully participated in TRECVID annually since 2005 [12, 13].

7.6 Semantic concept detection

Extracting semantic concepts from visual data has attracted a lot of research attention recently. The aim of the research has been to facilitate semantic indexing and concept-based retrieval of unannotated visual content. The leading principle has been to build semantic representations by obtaining intermediate semantic levels (objects, locations, events, activities, people, etc.) from automatically extracted low-level features. The modeling of mid-level semantic concepts can be useful in supporting high-level indexing and querying on multimedia data, as such concept models can be trained off-line with considerably more positive and negative examples than what are available at query time.

We treat semantic concept detection from shot-segmented videos as a general supervised classification task by utilizing the hierarchical approach shown in Fig. 7.4 and by extracting multiple low-level features from the different data modalities [14]. A set of SOMs is trained on these features to provide a common indexing structure across the different modalities. The particular features used for each concept detector are obtained using sequential forward feature selection. The method has proven to be readily scalable to a large number of concepts, which has enabled us to model e.g. a total of 294 concepts from a large-scale multimedia ontology [15] and utilize these concept models in TRECVID video search experiments [12]. Figure 7.5 lists and exemplifies the 36 semantic concepts detected for the TRECVID 2007 high-level feature extraction task.

Figure 7.5: The set of 36 semantic concepts used in TRECVID 2007.

Semantic concepts do not exist in isolation, but have different relationships between each other, including similarities in their semantic and visual (low-level) characteristics, co-occurrence statistics, and different hierarchical relations if a taxonomy has been defined for the concepts. We have studied how multimedia concept models built over a general clustering method can be interpreted in terms of probability distributions and how the quality of such models can be assessed with entropy-based methods [16].

In addition we also explored the possibility of taking advantage of temporal and inter-concept co-occurrence patterns of the high-level features using n-gram models and clustering of temporal neighborhoods. The method was found to be very useful in our TRECVID 2007 experiments [13].

7.7 Shot boundary detection

We have applied our general multimedia analysis framework to shot boundary detection and summarization of video data. Our approach for shot boundary detection utilizes the topology preservation properties of SOMs in spotting the abrupt and gradual shot transitions. Multiple feature vectors calculated from consecutive frames are projected on two-dimensional feature-specific SOMs. The transitions are detected by observing the trajectories formed on the maps.

Due to the topology preservation, similar inputs are mapped close to one another on the SOMs. The trajectory of the best-matching map units of successive frames thus typically hovers around some region of a SOM during a shot, provided that the visual content of the video does not change too rapidly. Abrupt cuts are characterized by sudden trajectory leaps from one region on the map to another, and gradual transitions on the other hand are characterized by a somewhat rapid drift of the trajectory from one region to another. The detector tries to detect these kinds of characteristic phenomena.

To increase detector robustness and prevent false positive cut detection decisions, e.g. due to flashlights, we do not only monitor the rate of change of the map position between two consecutive frames, but take small frame windows from both sides of the current point of interest, and compare the two frame windows. A circular area with a constant radius is placed over each map point in the given frame window as illustrated in Figure 7.6. We call the union of these circular areas the area spanned by the frame window. If the areas spanned by the preceding and following frame windows overlap, there are some similar frames in both of them, and we decide that the current point of interest is not a boundary point. If there is no overlapping, the frames in the frame windows are clearly dissimilar, and we decide that we have found a boundary. The flashlights are characterized by sudden trajectory leaps to some region on the map followed by a leap back to the original region. If the duration of the flashlight is smaller than the frame window size, the proposed method helps to avoid false positives.

The final boundary decision is done by a committee machine that consists of this kind of parallel classifiers. There is one classifier for each feature calculated from the frames, and each classifier has a weight value. The final decision is made by comparing the weighted vote result of the classifiers against a threshold value. Abrupt cuts and gradual transitions are detected using the same method. The detected boundary points that are close to one another are combined, and as the result we get the starting locations and lengths of the transitions. To facilitate detection of slow gradual transitions, our system also allows to use a frame gap of given length between the two frame windows. A more detailed description and quantitative results with the algorithm are given in [17].

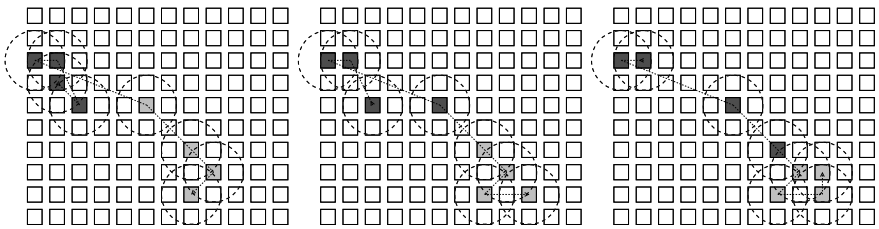


Figure 7.6: Segments of a trajectory at three consecutive time steps. The SOM cells marked with a dark gray color represent trajectory points belonging to the set of preceding frames, and light gray cells represent the following frames. The circles represent the area spanned by the preceding and following frame sets.

7.8 Video summarization

Video summarization is a process where an original video file is converted to a considerably shorter form. The video summary can then be used to facilitate efficient searching and browsing of video files in large video collections. The aim of successful automatic summarization is to preserve as much as possible from the essential content and overall structure. Straightforward methods such as frame subsampling and fast forwarding produce incoherent summaries that are strenuous to view and cannot usually be absorbed with a single viewing. The strategy of selecting parts of the video using a fixed interval can easily lose important information. More sophisticated summarization algorithms typically use shot-based segmentation and analysis. However, including each shot in the summary may not be optimal as certain shots may be almost duplicates of each other or there may be too many of them for a concise summary, depending on the original material.

There are two fundamental types of video summaries: *static abstracts or storyboards* and *video skims*. The former typically consist of collections of keyframes extracted from the video material and organized as a temporal timeline or as a two-dimensional display. Video skims consist of collections of selected video clips from the original material. Both these types of summaries can be useful, depending on the intended application. Storyboards provide static overviews that are easily presented and browsed in many environments, whereas skims preserve the original media type and can also contain dynamic content such as important events in the original video.

Figure 7.7: Representative frames and SOM signatures of three video shots.

We have developed a technique for video summarization as video skims [18] using SOMs trained with standard visual features that have been applied in various multimedia analysis tasks. The method is based on initial shot boundary detection providing us with lists of shots, which are used in the following stages as basic units of processing. We detect and remove unwanted "junk" shots (e.g. color bar test screens, empty frames) from the videos, and apply face detection and motion activity estimation. Next, we compute the visual similarities between all pairs of shots and remove overly similar shots. We trace the trajectory of the frames within the shot in question and record the corresponding BMUs. The set of BMUs constitutes a SOM-based signature for the shot, which can then be compared to other shots' signatures to determine whether a shot is visually unique or similar to some other shots. Fig. 7.7 shows example frames from three shots and the convolved SOM-based trajectory signatures of those shots as red-colored responses on the SOM surfaces. Each remaining shot is then represented in the summary with a separately selected one-second clip. The selected clips are finally combined using temporal ordering and fade-outs and fade-ins from black.

We participated in the TRECVID 2007 rushes summarization task [18] and obtained very promising results. Our summarization algorithm obtained average ground-truth inclusion performance with the shortest overall summaries over all the submissions.

References

[1] Mark Everingham, Andrew Zisserman, Chris Williams, and Luc Van Gool. The Pascal Visual Object Classes Challenge 2006 (VOC2006) results. Technical report, 2006. Available on-line at `http://www.pascal-network.org/`.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[3] Ville Viitaniemi and Jorma Laaksonen. Techniques for still image scene classification and object detection. In *Proceedings of 16th International Conference on Artificial Neural Networks (ICANN 2006)*, volume 2, pages 35–44, Athens, Greece, September 2006. Springer.

[4] Ville Viitaniemi and Jorma Laaksonen. Evaluating the performance in automatic image annotation: example case by adaptive fusion of global image features. *Signal Processing: Image Communications*, 22(6):557–568, July 2007.

[5] Ville Viitaniemi and Jorma Laaksonen. Improving the accuracy of global feature fusion based image categorisation. In Bianca Falcidieno, Michela Spagnuolo, Yannis S. Avrithis, Ioannis Kompatsiaris, and Paul Buitelaar, editors, *Proceedings of the 2nd International Conference on Semantic and Digital Media Technologies (SAMT 2007)*, volume 4669 of *Lecture Notes in Computer Science*, pages 1–14, Genova, Italy, December 2007. Springer.

[6] Zhirong Yang and Jorma Laaksonen. Interactive content-based facial image retrieval with partial relevance and parzen discriminant analysis. *Pattern Recognition Letters*, 2008. In submission.

[7] Zhirong Yang and Jorma Laaksonen. Face recognition using Parzenfaces. In *Proceedings of International Conference on Artificial Neural Networks (ICANN'07)*, volume 4669 of *Lecture Notes in Computer Science*, pages 200–209, Porto, Portugal, September 2007. Springer.

[8] Zhirong Yang and Jorma Laaksonen. Principal whitened gradient for information geometry. *Neural Networks*, 2008. In press.

[9] Matthieu Molinier, Jorma Laaksonen, and Tuomas Häme. Detecting man-made structures and changes in satellite imagery with a content-based information retrieval system built on self-organizing maps. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4):861–874, April 2007.

[10] Matthieu Molinier, Jorma Laaksonen, Yrjö Rauste, and Tuomas Häme. Detecting changes in polarimetric SAR data with content-based image retrieval. In *Proceedings of IEEE International Geoscience And Remote Sensing Symposium*, Barcelona, Spain, July 2007. IEEE.

[11] Erkki Oja, Mats Sjöberg, Ville Viitaniemi, and Jorma Laaksonen. Emergence of semantics from multimedia databases. In Gary Y. Yen and David B. Fogel, editors, *Computational Intelligence: Principles and Practice*, chapter 9. IEEE Computational Intelligence Society, 2006.

[12] Mats Sjöberg, Hannes Muurinen, Jorma Laaksonen, and Markus Koskela. PicSOM experiments in TRECVID 2006. In *Proceedings of the TRECVID 2006 Workshop*, Gaithersburg, MD, USA, November 2006.

[13] Markus Koskela, Mats Sjöberg, Ville Viitaniemi, Jorma Laaksonen, and Philip Prentis. PicSOM experiments in TRECVID 2007. In *Proceedings of the TRECVID 2007 Workshop*, Gaithersburg, MD, USA, November 2007.

[14] Markus Koskela and Jorma Laaksonen. Semantic concept detection from news videos with self-organizing maps. In Ilias Maglogiannis, Kostas Karpouzis, and Max Bramer, editors, *Proceedings of 3rd IFIP Conference on Artificial Intelligence Applications and Innovations*, pages 591–599, Athens, Greece, June 2006. IFIP, Springer.

[15] Milind Naphade, John R. Smith, Jelena Tešić, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[16] Markus Koskela, Alan F. Smeaton, and Jorma Laaksonen. Measuring concept similarities in multimedia ontologies: Analysis and evaluations. *IEEE Transactions on Multimedia*, 9(5):912–922, August 2007.

[17] Hannes Muurinen and Jorma Laaksonen. Video segmentation and shot boundary detection using self-organizing maps. In *Proceedings of 15th Scandinavian Conference on Image Analysis (SCIA 2007)*, pages 770–779, Aalborg, Denmark, June 2007.

[18] Markus Koskela, Mats Sjöberg, Jorma Laaksonen, Ville Viitaniemi, and Hannes Muurinen. Rushes summarization with self-organizing maps. In *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, pages 45–49, Augsburg, Germany, September 2007. ACM Press.

Chapter 8

Automatic speech recognition

Mikko Kurimo, Kalle Palomäki, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Ville Turunen, Sami Virpioja, Matti Varjokallio, Ulpu Remes, Antti Puurula

8.1 Introduction

Automatic speech recognition (ASR) means an automated process that inputs human speech and tries to find out what was said. ASR is useful, for example, in speech-to-text applications (dictation, meeting transcription, etc.), speech-controlled interfaces, search engines for large speech or video archives, and speech-to-speech translation.

Figure 8.1 illustrates the major modules of an ASR system and their relation to applications. In *feature extraction*, signal processing techniques are applied to the speech signal in order to dig out the features that distinguish different phonemes from each other. Given the features extracted from the speech, *acoustic modeling* provides probabilities for different phonemes at different time instants. *Language modeling*, on the other hand, defines what kind of phoneme and word sequences are possible in the target language or application at hand, and what are their probabilities. The acoustic models and language models are used in *decoding* for searching the recognition hypothesis that fits best to the models. Recognition output can then be used in various applications.

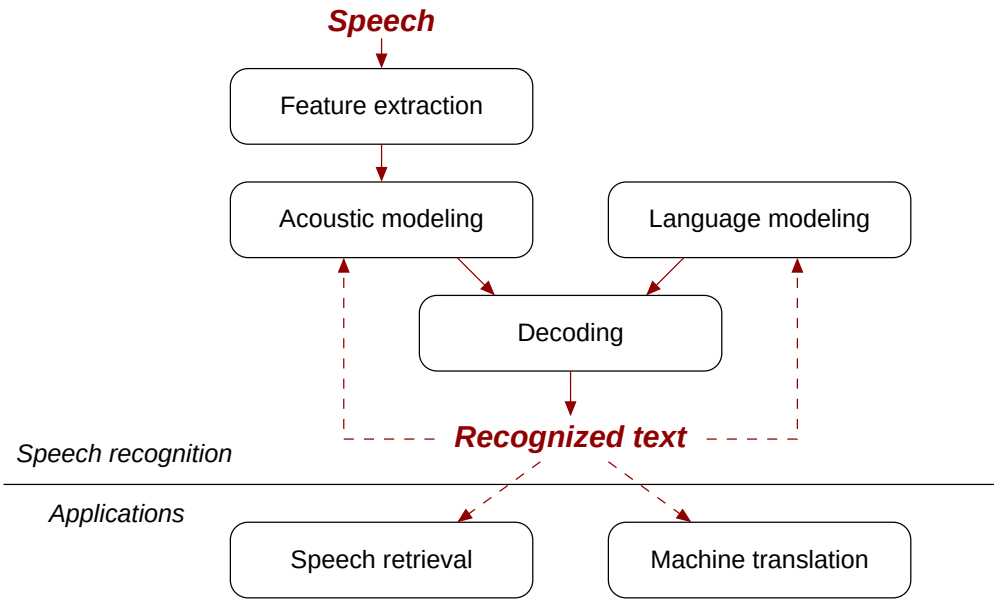


Figure 8.1: The main components of an automatic speech recognition system and their relation to speech retrieval and machine translation applications.

Our focus in ASR is large vocabulary continuous speech recognition (LVCSR). For several years, we have been developing new machine learning algorithms for each of the subfields and building a complete state-of-the-art recognizer to evaluate new methods and their impact. Originally, the recognizer was constructed for fluent and planned speech such as Finnish newsreading, where language models covering a very large vocabulary are required. Besides newsreading, other example tasks are political and academic speeches and other radio and television broadcasts where the language used is near the written style. Sofar, we have not seriously attempted to recognize Finnish spontaneous conversations, because enough Finnish training texts for learning the corresponding style do not exist. Our main training corpus for language modeling is the Finnish Language Bank at CSC. For acoustic modeling we use voice books, Finnish Broadcast Corpus at CSC and the SPEECON corpus.

In addition to the recognition of Finnish, we have performed experiments in English, Turkish and Estonian. To make this possible we have established research relations to

different top speech groups in Europe and U.S., e.g. University of Colorado, International Computer Science Institute ICSI, Stanford Research Institute SRI, IDIAP, University of Edinburgh, University of Sheffield, Bogazici University, and Tallinn University of Technology. The forms of collaboration have included researcher exchanges, special courses, workshops and joint research projects. We have also participated in several top international and national research projects funded by EU, Academy of Finland, Tekes, and our industrial partners. In the close collaboration with our Natural Language Processing group 10 we are also organizing an international competition called Morphochallenge to evaluate the best unsupervised segmentation algorithms for words into morphemes for information retrieval, LVCSR and language modeling in different languages. This challenge project is funded by EU's PASCAL network and described in Chapter 10.

8.2 Acoustic modeling

Acoustic modeling in automatic speech recognition means building statistical models for some meaningful speech units based on the feature vectors computed from speech. In most systems the speech signal is first chunked into overlapping 20-30 ms time windows at every 10 ms and the spectral representation is computed from each frame. A commonly used feature vector consists of mel-frequency cepstral coefficients (MFCC) which are the result of the discrete cosine transform (DCT) applied to the logarithmic mel-scaled filter bank energies. Local temporal dynamics can be captured by concatenating the first and second order delta features (time differences) to the basic feature vector.

Although the use of DCT and delta features are well-established methods for processing speech features, they are by no means optimal. Better features can be constructed by learning from the data which features would best discriminate between speech sounds. A well known method for this is the linear discriminant analysis (LDA), which can be used to process the spectral input for creating new discriminative features. As a simple method LDA has its limitations, and therefore in [1] we studied different methods to enhance its operation. The result was the pairwise linear discriminant (PLD) features, which unlike most LDA extensions are simple to compute but still work in speech recognition better than the traditional methods.

Closely connected to the feature extraction is the speaker-wise normalization of the features. One commonly used method is the vocal tract length normalization (VTLN). It requires estimating only a single normalization parameter yet still provides significant improvements to the speech recognition. The estimation, however, can not be done in closed form, so an exhaustive search over a range of parameters is usually used. We have devised a method which greatly simplifies the estimation of the VTLN parameter but still gives competitive performance [2]. It is especially attractive when used with discriminative feature extraction, such as with PLD.

The acoustic feature sequence in ASR is typically modeled using hidden Markov models (HMM). In basic form each phoneme is modeled by a separate HMM, where the emission distributions of the HMM states are Gaussian mixtures. An example is shown in Figure 8.2. In practice, however, we need to take the phoneme context into account, so that for each phoneme there are separate HMMs for various phoneme contexts. This leads easily to very complex acoustic models where the number of parameters is in order of millions.

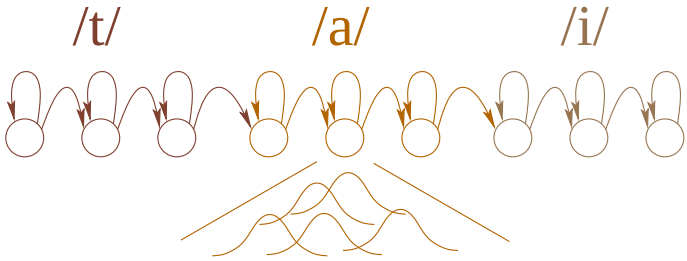


Figure 8.2: Each phoneme is modeled with a hidden Markov model, usually consisting of three states. The state distributions are modeled by Gaussian mixture models.

To limit the number of parameters and thereby allow robust estimation of the acoustic models, the covariance matrices of the Gaussian mixture components are usually assumed diagonal. This is a relatively reasonable assumption, because there is typically a whitening transform (DCT or similar) applied to the feature vector. The uncorrelatedness is, however, a global property and there are always correlations on the state level. The correlations can be modeled by adding more mixture components in the direction of most

variance, which is sometimes called as *implicit covariance modeling*. Modeling covariances *explicitly* instead has some clear benefits as fewer modeling assumptions typically lead to more robust models. Constraining the exponential parameters of the Gaussians to a subspace is appealing for speech recognition, as the computational cost of the acoustic model is also decreased. A subspace constraint on the inverse covariance matrices was shown to give a good performance [3] for LVCSR tasks.

To ensure high quality research we constantly put considerable effort to keep our speech recognition system up-to-date. One major recent improvement to our system has been the introduction of discriminative acoustic training. The use of discriminative training has been a growing trend during the last decade and some form of it is now a necessity for a state-of-the-art system. Our implementation allows using several different training criteria such as maximum mutual information (MMI) and minimum phone error (MPE) [4] over the traditional maximum likelihood (ML) training. It also enables gradient based optimization in addition to the commonly used extended Baum-Welch method. Discriminative training techniques have already given very promising results and they will be an important research direction in the future.

Speaker segmentation

In addition to feature normalization methods such as the vocal tract length normalization (VTLN), acoustic model adaptation is often used for increased robustness against speaker variation. Speaker normalization and adaptation generally improve the speech recognition performance substantially, but they cannot be applied unless the speech recognition system knows who spoke and when. Often there is no such information about the speakers, but automatic speaker segmentation is needed. Speaker segmentation (i) divides the audio to speaker turns (speaker change detection) and (ii) labels the turns according to speaker (speaker tracking) as illustrated in Figure 8.3. While most speaker segmentation methods have been developed primarily for audio content or spoken dialogue analysis, we focused on speaker segmentation for speaker adaptation. We developed a speaker tracking method that seeks to directly maximize the feature likelihood when we assume the features are adapted to speaker using the segmentation results and acoustic model adaptation with constrained maximum likelihood linear regression (CMLLR). The proposed method performed well when tested on Finnish television news audio in [5].

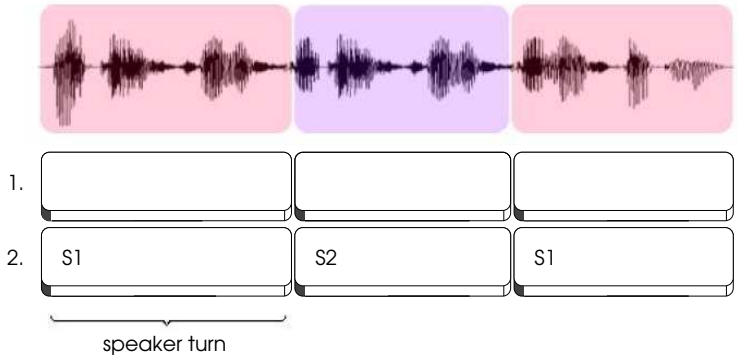


Figure 8.3: Speaker segmentation first divides the audio to speaker turns according to where speakers change and then labels the detected turns. Speaker labels are created on-line and no prior information about the speakers (e.g. training data or speaker models) is needed.

Recognition of reverberant speech

Research in the acoustic modeling for large vocabulary continuous speech recognition was concentrated mostly on fairly noise free conditions (see Sect. 8.2). In the field noise robust speech recognition we have been developing techniques suitable for recognition in highly reverberant spaces. This research has been collaborative with the University of Sheffield. Our approach is based on missing data approach [9], in which noisy, reverberated regions are treated as unreliable and noise free regions as reliable evidence of speech. Different treatments of reliable and unreliable parts of speech is achieved by a modification of Gaussian mixture model proposed by Cooke et al. [9]. Our approach to reverberant speech recognition is based on detecting reliable regions of speech from strong onsets at modulation rates characteristic to speech [8]. In recent developments of the model we have sought modeling solutions that more closely match on perceptual data considering the recognition of reverberant speech by human listeners [6, 7].

References

[1] J. Pylkkönen, LDA Based Feature Estimation Methods for LVCSR. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh PA, USA, pages 389–392, 2006.

[2] J. Pylkkönen, Estimating VTLN Warping Factors by Distribution Matching. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Antwerp, Belgium, pages 270–273, 2007.

[3] M. Varjokallio, M. Kurimo, Comparison of Subspace Methods for Gaussian Mixture Models in Automatic Speech Recognition. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Antwerp, Belgium, pages 2121-2124, 2007.

[4] D. Povey and P. C. Woodland, Minimum Phone Error and I-smoothing for Improved Discriminative Training. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, pages I-105–108, 2002.

[5] U. Remes, J. Pylkkönen, and M. Kurimo, Segregation of Speakers for Speaker Adaptation in TV News Audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, USA, pages IV-481–484, 2007.

[6] G. J. Brown and K. J. Palomäki Reverberation, in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, eds. by DeLiang Wang and Guy J. Brown, Wiley/IEEE Press, 2006.

[7] G. J. Brown and K. J. Palomäki A reverberation-robust automatic speech recognition system based on temporal masking, Research abstract accepted to Acoustics 2008, Paris, France.

[8] K. J. Palomäki, G. J. Brown and J. Barker, Recognition of reverberant speech using full cepstral features and spectral missing data,*Proceedings the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tolouse, France, vol. 1, 289-292, 2006.*

[9] M.P. Cooke, P. Green, L. Josifovski, and A. Vizinho, Robust automatic speech recognition with missing and unreliable acoustic data,*Speech Comm.*, vol. 34, pp. 267 285, 2001.

8.3 Language modeling

For Finnish, estimating the language model probabilities for words is difficult since there is a vast number of different word forms. For example, a single verb has theoretically thousands of inflected word forms. The natural way to attack the problem is to split words into smaller fragments and build the language models on the fragments instead of whole words. Since it is not obvious how the words should be split, we have studied what kind of word fragments are optimal for speech recognition systems. Experiments in Finnish, Turkish and Estonian recognition tasks indicate that an unsupervised data-driven splitting algorithm called Morfessor (see Section 10.1) improves recognition of rare words. [1]

N-gram models are the most widely used language models in large vocabulary continuous speech recognition. Since the size of the model grows rapidly with respect to the model order and available training data, many methods have been proposed for pruning the least relevant n-grams from the model. However, correct smoothing of the n-gram probability distributions is important and performance may degrade significantly if pruning conflicts with smoothing. In the journal paper [2] we show that some of the commonly used pruning methods do not take into account how removing an n-gram should modify the backoff distributions in the state-of-the-art Kneser-Ney smoothing. We also present two new algorithms: one for pruning Kneser-Ney smoothed models, and one for growing them incrementally. Experiments on Finnish and English text corpora show that the proposed pruning algorithm provides considerable improvements over previous pruning algorithms on Kneser-Ney smoothed models and is also better than the baseline entropy pruned Good-Turing smoothed models.

Representing the language model compactly is important in recognition systems targeted for small devices with limited memory resources. In [3], we have extended the compressed language model structure proposed earlier in the literature. By separating n-grams that are prefixes to longer n-grams, redundant information can be omitted. Experiments on English 4-gram models and Finnish 6-gram models show that extended structure can achieve up to 30 % lossless memory reductions when compared to the baseline structure.

Another common method for decreasing the size of the n-gram models is clustering of the model units. However, if size of the lexicon is very small, as in models based on statistical morpheme-like units (see, e.g., [1]), clustering of individual units is not so useful. Instead, we have studied how sequences of the morpheme-like units can be clustered to achieve improvements in speech recognition. When the clustered sequences are histories (context parts) of the n-grams, it is easy to combine the clustering to the incremental growing of the model applied in, e.g., [2]. Maximum a posteriori estimation can be used to make a compromise between the model size and accuracy. The experiments show that the clustering is useful especially if very compact models are required. [4]

References

[1] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraçlar, and A. Stolcke. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing*, 5(1), 2007.

[2] V. Siivola, T. Hirsimäki, and S. Virpioja. On Growing and Pruning Kneser-Ney Smoothed N-Gram Models. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5), pages 1617–1624, 2007.

[3] T. Hirsimäki. On Compressing N-gram Language Models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, USA, pages IV-949–952, 2007.

[4] S. Virpioja and M. Kurimo. Compact N-gram Models by Incremental Growing and Clustering of Histories. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh PA, USA, pages 1037–1040, 2006.

8.4 Applications and tasks

Speech retrieval and indexing

Large amounts of information is produced in spoken form. In addition to TV and radio broadcasts, more and more material is distributed on the Internet in the form of podcasts and video sharing web sites. There is an increasing need for content based retrieval of this material. Speech retrieval systems consist of two parts as illustrated in Figure 8.4. First, an automatic speech recognition system is used to transcribe the speech into textual form. Second, an index is built based on this information.

The vocabulary of the speech recognizer limits the possible words that can be retrieved. Any word that is not in the vocabulary will not be recognized correctly and thus can not be used in retrieval. This is especially problematic since the rare words, such as proper names, that may not be in the vocabulary are often the most interesting from retrieval point of view. Our speech retrieval system addresses this problem by using morpheme-like units produced by the Morfessor algorithm. Any word in speech can now potentially be recognized by recognizing its component morphemes. The recognizer transcribes the text as a string of morpheme-like units and these units can also be used as index terms.

One problem of using morpheme-like units as index terms is that different inflected forms of the same word can produce different stems when they are split to morphemes. However, we would like to retrieve the speech document no matter what inflected form of the word is used. This resembles the problem of synonyms. We have countered this problem by applying Latent Semantic Indexing to the morpheme-based retrieval approach [1]. The method projects different stems of the same word to the same dimension that represents the true, latent, meaning of the term.

Speech recognizers typically produce only the most likely string of words, the 1-best hypothesis. Retrieval performance is decreased if a relevant term is misrecognized and is thus missing from the transcript. However, it is possible that the correct term was considered by the recognizer but was not the top choice. Thus, retrieval performance can be improved by extracting these alternative results from the recognizer and adding them to the index. A *confusion network* [2] provides a convenient representation of the competing terms along with a probability value for each term. However, as most terms in the network were in fact not spoken, the indexing method must be designed so that it is not degraded by these spurious terms. In [3], we compare methods that use the probability and rank of the terms to weigh the index terms properly and show improved performance of the retrieval system.

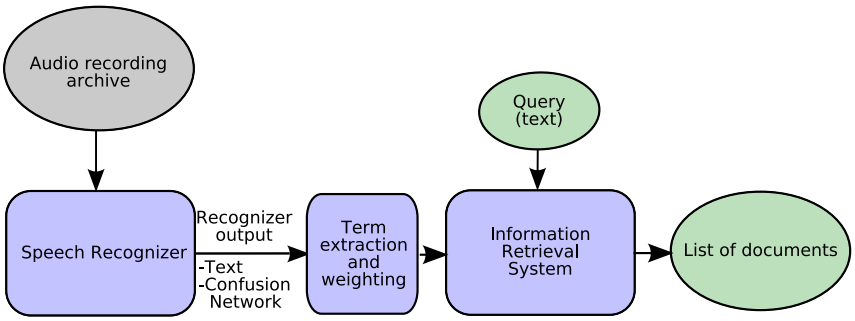


Figure 8.4: Overview of a spoken document retrieval system.

Estonian speech recognition

For agglutinative languages, like Finnish, Estonian, Hungarian and Turkish, it is practically impossible to build a word-based lexicon for speech recognition that would cover all the relevant words. The problem is that words are generally formed by concatenating several prefixes and suffixes to the word roots. Together with compounding and inflections this leads to millions of different, but still frequent word forms that can not be trivially split into meaningful parts. For some languages there exists rule-based morphological analyzers that can perform this splitting, but they are laborious to create and due to the handcrafted rules, they also suffer from an out-of-vocabulary problem.

In a pilot study of language and task portability of our speech recognition and language modeling tools, we created an Estonian speech recognizer. The text corpus used to learn the morph units and train the statistical language model consisted of newspapers and books, altogether about 55 million words [4]. The speech corpus consisted of over 200 hours and 1300 speakers, recorded from telephone [5], i.e. 8 kHz sampling rate and narrow band data instead of 16 kHz and normal (full) bandwidth that we have used for Finnish data. The speaker independence, together with the telephone quality and occasional background noises, made this task more difficult than our Finnish ones, but with the help of our learning and adaptive models we were still able to reach good recognition results and demonstrate a performance that was superior to the word-based reference systems [6, 7].

Speech-to-speech translation

Speech-to-speech machine translation is in some ways the peak of natural language processing, in that it deals directly with our (humans') original, oral mode of communication (as opposed to derived written language). As such, it presents several important challenges:

1. Automatic speech recognition of the input using state-of-the-art acoustic and language modeling, adaptation and decoding
2. Statistical machine translation of either the recognized most likely speech transcript or the confusion network or the whole lattice including all the best hypothesis
3. Speech synthesis to turn the translation output into intelligible speech using the state-of-the-art synthesis models and adaptation
4. Intergration of all these components to aim at the best possible output and tolerate errors that may happen in each phase

A pilot study of Finnish-English speech-to-speech translation was carried out in the lab as a joint effort of the speech recognition, Natural Language Processing 10 and Computational Cognitive Systems 10.3 groups. The domain selected for our experiments was heavily influenced by the available bilingual (Finnish and English) and bimodal (text and speech) data. Because none is readily yet available, we put one together using the Bible. As the first approach we utilized the existing components, and tried to weave them together in an optimal way. To recognize speech into word sequences we applied our morpheme-based unlimited vocabulary continuous speech recognizer [8]. As a Finnish acoustic model the system utilized multi-speaker hidden Markov models with Gaussian mixtures of mel-cepstral input features for state-tied cross-word triphones. The statistical language model was trained using our growing varigram model [9] with unsupervised morpheme-like units derived from Morfessor Baseline [10]. In addition to the Bible the training data included texts from various sources including newspapers, books and newswire stories totally about

150 million words. For translation, we trained the Moses system [11] on the same word and morpheme units as utilized in the language modeling units of our speech recognizer. For speech synthesis, we used Festival [12], including the built-in English voice and a Finnish voice developed at University of Helsinki.

References

[1] V. Turunen and M. Kurimo Using Latent Semantic Indexing for Morph-based Spoken Document Retrieval, *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh PA, USA, pages 389–392, 2006.

[2] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech And Language*, 14:373–400, 2000.

[3] V. Turunen and M. Kurimo Indexing Confusion Networks for Morph-based Spoken Document Retrieval, *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and development in information retrieval*, Amsterdam, The Netherlands, pages 631–638, 2007.

[4] Segakorpus. 2005. Segakorpus - Mixed Corpus of Estonian. Tartu University. *http://test.cl.ut.ee/korpused/*.

[5] Einar Meister, Jürgen Lasn and Lya Meister 2002. Estonian SpeechDat: a project in progress. In *Proceedings of the Fonetiikan Päivät - Phonetics Symposium 2002 in Finland*, 21–26.

[6] Mikko Kurimo, Antti Puurula, Ebru Arisoy, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Tanel Alumae and Murat Saraclar 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics*. HLT-NAACL 2006. New York, USA

[7] Antti Puurula and Mikko Kurimo 2007. Vocabulary Decomposition for Estonian Open Vocabulary Speech Recognition. In *Proceedings of ACL 2007*.

[8] Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja and Janne Pylkkönen 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language* 20(4):515–541.

[9] Vesa Siivola Language models for automatic speech recognition: construction and complexity control. Doctoral thesis, Dissertations in Computer and Information Science, Report D21, Helsinki University of Technology, Espoo, Finland, 2006.

[10] Mathias Creutz. Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition. Doctoral thesis, Dissertations in Computer and Information Science, Report D13, Helsinki University of Technology, Espoo, Finland, 2006.

[11] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Ondrej Bojar, Alexandra Constantin, and Evan Herb. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, 2007.

[12] The Festival Speech Synthesis System. University of Edinburgh. *http://festvox.org*

Chapter 9

Proactive information retrieval

Samuel Kaski, Kai Puolamäki, Antti Ajanki, Jarkko Salojärvi

9.1 Introduction

Successful proactivity, that is anticipation, in varying contexts requires generalization from past experience. Generalization, on its part, requires suitable powerful (stochastic) models and a collection of data about relevant past history to learn the models.

The goal of the PROACT project is to build statistical machine learning models that learn from the actions of people to model their intentions and actions. The models are used for disambiguating the users' vague commands and anticipating their actions.

Our application area is information retrieval, where we investigate to what extent the laborious explicit relevance feedback can be complemented or even replaced by implicit feedback derived from patterns of eye fixations and movements that exhibit both voluntary and involuntary signs of users intentions. Inference is supported by models of document collections and interest patterns of users.

The PROACT project has been done in close collaboration with researchers in the European Union's Pascal Network of Excellence within a Pump Priming Programme (2005–2007); the collaborators are from University of Helsinki, University of Southampton and University College London. The project continues in a STREP project PinView from 2008 onwards.

9.2 Implicit queries from eye movements

Eye movements measured during reading are a promising new source of implicit feedback. During complex tasks such as reading, attention approximately lies on the location of the reader's gaze. Therefore the eye movements should contain information on the reader's interests. Inferring interest of the user from a reading pattern is difficult however, since the signal is complex and very noisy, and since interestingness or relevance is higly subjective and thus hard to define. We have earlier developed machine learning and signal processing methods for this task, and hosted a research challenge where the task was to predict relevance from eye movement patterns [1].

The motivation for the next stage of the research was that formulating a good query in a web search engine, for example, is known to be difficult. Implicit feedback collected by observing user's behavior might reveal the true interest of the user without the need to explicitely label the documents as relevant or not relevant. We performed a feasibility study in which we used eye movements to formulate a query that reflects user's interest while he was reading [2].

We constructed a controlled experimental setting in which it is known which documents are relevant. The users read short documents searching for the ones related to a topic that was given to them beforehand. The eye movements were recorded during reading. We trained a regressor that predicts how relevant a term is for user's current query given the eye movement measurements on that term. This regressor can then be applied to new topics, with no training data available, to estimate relevance of words.

The learned model was then used to infer relevant query terms based on eye movement recorded while the user was performing a new search task, where the true query was unknown. The inferred query terms can be used to retrieve and suggest new documents that might be important to user's information need.

A SVM model that uses eye movements and textual features outperformed a similar model without the eye movement features. This indicates that eye movements contain exploitable information about relevance in information retrieval tasks.

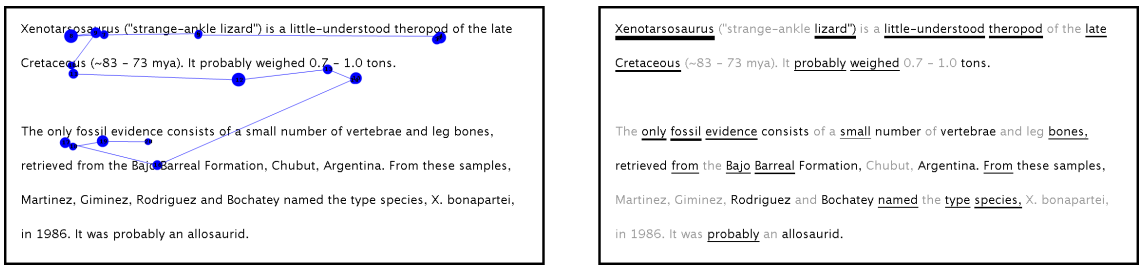


Figure 9.1: Left: A sample eye movement pattern of a test subject during reading a document. Right: The term weights depicted on the same document inferred from the eye movements of all test subjects who were searching for information about dinosaurs. The magnitudes of the weights are depicted as the thickness of the underlining.

References

[1] Kai Puolamäki and Samuel Kaski, editors. *Proceedings of the NIPS 2005 Workshop on Machine Learning for Implicit Feedback and User Modeling*. Helsinki University of Technology, Espoo, Finland, 2006.

[2] David R. Hardoon, John Shawe-Taylor, Antti Ajanki, Kai Puolamäki, and Samuel Kaski. Information Retrieval by Inferring Implicit Queries from Eye Movements. In *Proceedings of the 11th International Conference on International Conference on Artificial Intelligence and Statistics*. San Juan, Puerto Rico, 2007.

Chapter 10

Natural language processing

Krista Lagus, Mikko Kurimo, Timo Honkela, Mathias Creutz, Jaakko J. Väyrynen, Sami Virpioja, Ville Turunen, Matti Varjokallio

10.1 Unsupervised segmentation of words into morphs

In the theory of linguistic morphology, morphemes are considered to be the smallest meaning-bearing elements of language, and they can be defined in a language-independent manner. It seems that even approximative automated morphological analysis is beneficial for many natural language applications dealing with large vocabularies, such as speech recognition and machine translation. These applications usually make use of *words* as vocabulary units. However, for highly-inflecting and agglutinative languages, this leads to very sparse data, as the number of possible word forms is very high. Figure 10.2 shows the very different rates at which the vocabulary grows in various text corpora of the same size. The number of different unique word forms in the Finnish corpus is considerably higher than in the English ones, for example.

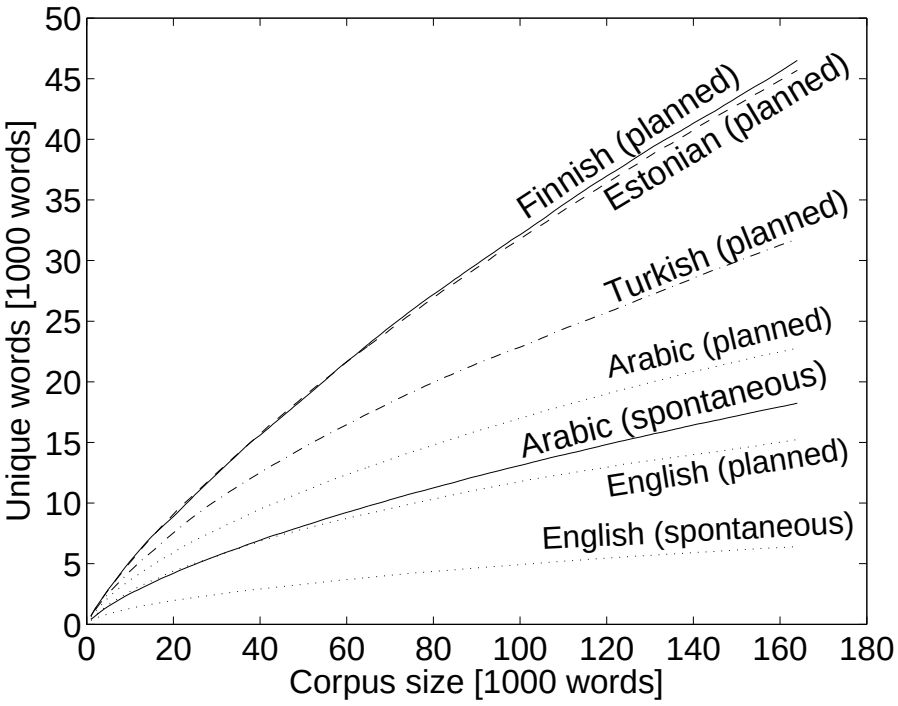


Figure 10.2: The number of different word forms (types) encountered in growing portions of running text (tokens) of various languages.

We have developed *Morfessor*, a language-independent, data-driven method for the unsupervised segmentation of words into morpheme-like units. There are different versions of Morfessor, which correspond to consecutive steps in the development of the model [1, 2, 3, 4]. All versions can be seen as instances of a general model, as described in [5].

The general idea behind the Morfessor model is to discover as compact a description of the data as possible. Substrings occurring frequently enough in several different word forms are proposed as *morphs* and the words are then represented as a concatenation of morphs, e.g., "hand, hand+s, left+hand+ed, hand+ful".

An optimal balance is sought between compactness of the *morph lexicon* versus the compactness of the representation of the *corpus*. The morph lexicon is a list of all distinct morphs (e.g., "hand, s, left, ed, ful") together with some stored properties of these morphs. The representation of the corpus can be seen as a sequence of pointers to entries in the morph lexicon; e.g. the word "lefthanded" is represented as three pointers to morphs in

the lexicon.

Among others, de Marcken [6], Brent [7], and Goldsmith [8] have shown that the above type of model produces segmentations that resemble linguistic morpheme segmentations, when formulated mathematically in a probabilistic framework or equivalently using the Minimum Description Length (MDL) principle [9].

A shortcoming of previous splitting methods is that they either do not model *context-dependency* or they *limit the number of splits* per word to two or three. Failure to incorporate context-dependency in the model may produce splits like "s+wing, ed+ward, s+urge+on" on English data, since the morphs "-s" and "-ed" are frequently occurring suffixes in the English language, but the algorithm does not make this distinction and thus suggests them in word-initial position as prefixes. By limiting the number of allowed segments per word the search task is alleviated and context-dependency can be modeled. However, this makes it impossible to correctly segment compound words with several affixes (pre- or suffixes), such as the Finnish word "aka+n+kanto+kiso+i+ssa" (transl. "in the wife-carrying contests").

We have focused our efforts on developing a segmentation model that incorporates context-dependency without restricting the number of allowed segments per word. This has resulted in two model variants, Categories-ML [3] and Categories-MAP [4]. The former is based on Maximum Likelihood (ML) optimization, in combination with some heuristics, whereas the latter applies a more elegant model formulation within the Maximum a Posteriori (MAP) framework. The MAP formulation, along with a thorough comparison to the other Morfessor variants, is provided also in [5] and [10].

Some sample segmentations of Finnish, English, as well as Swedish words, are shown in Figure 10.3. These include correctly segmented words, where each boundary coincides with a linguistic morpheme boundary (e.g., "aarre+kammio+i+ssa, edes+autta+isi+vat, abandon+ed, long+fellow+'s, in+lopp+et+s"). In addition, some words are over-segmented, with boundaries inserted at incorrect locations (e.g., "in+lägg+n+ing+ar" instead of "in+lägg+ning+ar"), as well as under-segmented words, where some boundary is missing (e.g., "bahama+saari+lla" instead of "bahama+saar+i+lla").

In addition to segmenting words, Morfessor suggests likely grammatical categories for the segments. Each morph is tagged as a prefix, stem, or suffix. Sometimes the morph categories can resolve the semantic ambiguity of a morph, e.g., Finnish "pää". In Figure 10.3, "pää" has been tagged as a stem in the word "pää+hän" ("in [the] *head*"), whereas it functions as a prefix in "pää+aihe+e+sta" ("about [the] *main* topic").

Evaluation

In the publications related to the development of Morfessor, the algorithm has been evaluated by comparing the results to linguistic morpheme segmentations of Finnish and English words [1, 2, 3, 4, 5]. In order to carry out the evaluation, linguistic reference segmentations needed to be produced as part of the project, since no available resources were applicable as such. This work resulted in a morphological "gold standard", called *Hutmegs* (Helsinki University of Technology Morphological Evaluation Gold Standard) [11, 12]. When the latest context-sensitive Morfessor versions [3, 4] are evaluated against the Hutmegs gold standard, they clearly outperform a frequently used benchmark algorithm [8] on Finnish data, and perform as well or better than the benchmark on English data.

Morfessor algorithms have also been evaluated in the Morpho Challenge competitions described in Section 10.2. Morpho Challenge 2007 included evaluation in four languages (English, Finnish, German and Turkish) and two competitions: comparison against linguistic standards and evaluation in information retrieval tasks. Morfessor managed fairly

aarre + **kammio** + *i* + *ssa*, **aarre** + **kammio** + *nsa*, **bahama** + **saar** + *et*, **bahama** + **saari** + *lla*, **bahama** + **saar** + *ten*, **edes** + **autta** + *isi* + *vat*, **edes** + **autta** + *ma* + *ssa*, nais + **auto** + *ili* + *ja* + *a*, pää + **aihe** + *e* + *sta*, pää + **aihe** + *i* + *sta*, **pää** + *hän*, taka + **penkki** + *lä* + *in*+ *en*, **voi** + *mme* + *ko*
abandon + *ed*, **abandon** + *ing*, **abandon** + *ment*, **beauti** + *ful*, **beauty** + *'s*, **calculat** + *ed*, **calculat** + *ion* + *s*, **express** + *ion* + *ist*, **micro** + **organ** + *ism* + *s*, **long** + **fellow** + *'s*, **master** + **piece** + *s*, **near** + *ly*, **photograph** + *er* + *s*, **phrase** + *d*, un + **expect** + *ed* + *ly*
ansvar + *ade*, **ansvar** + *ig*, **ansvar** + *iga*, **ansvar** + *s* + för + **säkring** + *ar*, **blixt** + ned + **slag**, **dröm** + *de*, **dröm** + *des*, **drömma** + *nde*, in + **lopp** + *et* + *s*, in + **lägg** + *n* + *ing* + *ar*, **målar** + *e*, **målar** + **yrke** + *t* + *s*, o + ut + **nyttja** + *t*, **poli** + *s* + **förening** + *ar* + *na* + *s*, **trafik** + **säker** + *het*, över + **fyll** + *d* + *a*

Figure 10.3: Examples of segmentations learned from data sets of Finnish, English, and Swedish text. Suggested prefixes are underlined, stems are rendered in **boldface**, and suffixes are *slanted*.

well in all the evaluations, especially with Finnish and Turkish languages.

Applications

Morfessor has been extensively tested as a component of a large vocabulary speech recognition system. By allowing a compact but flexible vocabulary for the system, Morfessor improves especially recognition of rare words. For several languages such as Finnish, Estonian and Turkish, this approach outperforms the state-of-the-art solutions. The speech recognition experiments are described in Section 8.3.

In addition to speech recognition, Morfessor has been used in speech retrieval and statistical machine translation systems. These experiments are described in Section 8.4 and 13, respectively.

Demonstration and software

There is an online demonstration of Morfessor on the Internet: `http://www.cis.hut.fi/projects/morpho/`. Currently, the demo supports three languages (Finnish, English, and Swedish) and two versions of the Morfessor (Baseline and Categories-ML). Those interested in larger-scale experiments can download the Morfessor program and train models using their own data sets. Two versions are available: Morfessor 1.0 software implements the Morfessor Baseline algorithm described in [13] and Morfessor Categories-MAP 0.9.2 software implements the Morfessor Categories-MAP algorithm described in [4]. During 2007, a monthly average of 10 downloads has been registered for both versions.

References

[1] Mathias Creutz and Krista Lagus. Unsupervised discovery of morphemes. In *Proc. Workshop on Morphological and Phonological Learning of ACL'02*, pages 21–30, Philadelphia, Pennsylvania, USA, 2002.

[2] Mathias Creutz. Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proc. ACL'03*, pages 280–287, Sapporo, Japan, 2003.

[3] Mathias Creutz and Krista Lagus. Induction of a simple morphology for highly-inflecting languages. In *Proc. 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51, Barcelona, July 2004.

[4] Mathias Creutz and Krista Lagus. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, 2005.

[5] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, Volume 4, Issue 1, Article 3, January 2007.

[6] C. G. de Marcken. *Unsupervised Language Acquisition*. PhD thesis, MIT, 1996.

[7] M. R. Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105, 1999.

[8] John Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, 2001.

[9] Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15. World Scientific Series in Computer Science, Singapore, 1989.

[10] Mathias Creutz. Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition. Doctoral thesis, Dissertations in Computer and Information Science, Report D13, Helsinki University of Technology, Espoo, Finland, 2006.

[11] Mathias Creutz and Krister Lindén. Morpheme segmentation gold standards for Finnish and English. Technical Report A77, Publications in Computer and Information Science, Helsinki University of Technology, 2004.

[12] Mathias Creutz, Krista Lagus, Krister Lindén, and Sami Virpioja. Morfessor and Hutmegs: Unsupervised morpheme segmentation for highly-inflecting and compounding languages. In *Proceedings of the Second Baltic Conference on Human Language Technologies*, pages 107–112, Tallinn, Estonia, 4 – 5 April 2005.

[13] Mathias Creutz and Krista Lagus. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology, 2005.

10.2 Morpho Challenge

Morpho Challenge is a series of scientific competition organized by Adaptive Informatics Research Centre for an evaluation of unsupervised morpheme analysis algorithms. The challenge is part of the EU Network of Excellence PASCAL Challenge Program and in 2007 organized in collaboration with Cross-Language Evaluation Forum CLEF. The objective of the challenge is to design statistical machine learning algorithms that discover which morphemes (smallest individually meaningful units of language) words consist of. Ideally, these are basic vocabulary units suitable for different tasks, such as text understanding, machine translation, information retrieval, and statistical language modeling. The challenge has sofar been organized two times: the results of the 2005 challenge were published in a workshop in April 2006 in Venice, Italy [1]. The 2007 challenge workshop was held in September 2007 in Budapest, Hungary [2, 3].

In the original challenge, the words were segmented in unsupervised morphemes and the results were evaluated by a comparison to linguistic gold standard morphemes. The organizers also used the results to for training statistical language models and evaluated the models in large vocabulary speech recognition experiments [1]. The 2007 challenge was a more difficult one requiring morpheme analysis of words instead of just segmentations into smaller units. The evaluation of the submissions was performed by two complementary ways: *Competition 1*: The proposed morpheme analyses were compared to a linguistic morpheme analysis gold standard by matching the morpheme sharing word pairs [2]. *Competition 2*: Information retrieval (IR) experiments were performed, where the words in the documents and queries were replaced by their proposed morpheme representations and the search was based on morphemes instead of words [3]. The IR evaluations were provided for Finnish, German, and English and participants were encouraged to apply their algorithm to all of them. The organizers performed the IR experiments using the queries, texts, and relevance judgments available in CLEF forum and morpheme analysis methods submitted by the challenge participants. The results show that the morpheme analysis has a significant effect in IR performance in all languages, and that the performance of the best unsupervised methods can be superior to the supervised reference methods.

References

[1] Mikko Kurimo, Mathias Creutz, Matti Varjokallio, Ebru Arisoy and Murat Saraclar. Unsupervised segmentation of words into morphemes - Challenge 2005, An Introduction and Evaluation Report. In *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*. Venice, Italy, April 12, 2006.

[2] Mikko Kurimo, Mathias Creutz, Matti Varjokallio. Unsupervised Morpheme Analysis Evaluation by a Comparison to a Linguistic Gold Standard – Morpho Challenge 2007. In *Working Notes of the CLEF 2007 Workshop.* Edited by Alessandro Nardi and Carol Peters. 19-21 September, Budapest, Hungary.

[3] Mikko Kurimo, Mathias Creutz, Ville Turunen. Unsupervised Morpheme Analysis Evaluation by IR experiments – Morpho Challenge 2007. In *Working Notes of the CLEF 2007 Workshop.* Edited by Alessandro Nardi and Carol Peters. 19-21 September, Budapest, Hungary.

10.3 Emergence of linguistic features using independent component analysis

We have been able to show that Independent Component Analysis (ICA) [1] applied on word context data provides distinct features that reflect syntactic and semantic categories [2]. The difference to latent semantic analysis (LSA) is that the analysis finds features or categories that are not only explicit but can also easily be interpreted by humans. This result can be obtained without any human supervision or tagged corpora that would have some predetermined morphological, syntactic or semantic information.

It is important to compare the capability of single features or feature pairs to separate categories because this measures how well the obtained features correspond with the categories. In fact, when all features are used, the separation capabilities of ICA and LSA are comparable because the total information present is the same. We have also shown that the emergent features match well with categories determined by linguists by comparing the ICA results to linguistic word category information [3].

We have shown how the features found by the ICA method can be further processed by simple nonlinear methods, such as thresholding, that gives rise to a sparse feature representation of words [4, 5]. We performed thresholding for each found word feature vector separately. The values closest to zero were set to zero and only a selected number of features were left to their original values. An analogical approach can be found from the analysis of natural images, where a soft thresholding of sparse coding is a denoising operator.

We compared the original representation and the thresholded representations in multiple choice vocabulary tasks, which measure the semantic information captured by the representation. An illustrative result is shown in Figure 10.4, which compares the feature thresholding with the two methods, latent semantic analysis and independent component analysis. The graph shows that the thresholded ICA representation is able to capture the most important semantics with fewer components, as the quality of the thresholded ICA representation degrades more slowly than both LSA representations. Several tests were run with three languages, including two different corpora, with quite similar results.

We have also shown how independent component analysis gives rise to a multilingual word feature space when trained with a parallel corpus [6]. The feature space created by the found features is also multilingual. Words that are related in different languages appear close to each other in the feature space, which makes it possible to find translations for words between languages. Table 10.1 shows the closest words for the English word 'finland' in the feature space, which include different forms of the Finnish equivalent, but also the name of a neighboring country ('sweden') as well as Austria ('itävalta'). The latter might be caused by shared work during the Finnish EU presidency. The single features also carry multilingual semantic information, as can be seen from Table 10.2, that lists the most prominent words in three features.

The attained results include both an emergence of clear distinctive categories or features and a distributed representation. In the emergent representation, a word may thus belong to several categories simultaneously in a graded manner. We see that further processing of the features is possible and thresholding produces a more sparse representation that can have greater interpretability without too much information loss. The method is also applicable to multilingual textual data, and is able to find representations where the multilingual semantic space can be used to mine translations and related words.

We wish that our model provides additional understanding on potential cognitive mechanisms in natural language learning and understanding Our approach attempts to show that it is possible that much of the linguistic knowledge is emergent in nature and based

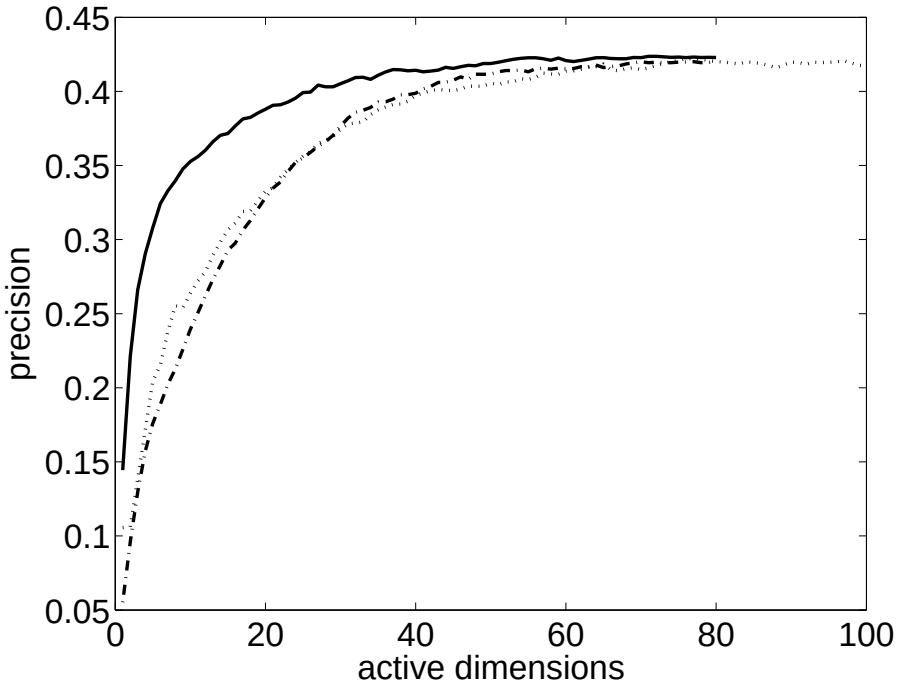


Figure 10.4: Rates of correctly answered questions with unthresholded LSA (dotted), LSA with thresholding with 80 components (dashed) and ICA with thresholding with 80 components (solid) set w.r.t. the number of non-zero features (after thresholding). The features where calculated from free electronic English books extracted from the Gutenberg project. The test questions were based on synonyms and related words extracted from the Moby thesaurus.

Table 10.1: The closest words in the multilingual feature space to the word 'finland'.

word	match
finland	1.00
suomen	0.83
suomi	0.82
sweden	0.79
suomessa	0.77
austria	0.73
...	...

on specific learning mechanisms.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis.* John Wiley & Sons, 2001.

[2] T. Honkela, and A. Hyvärinen. Linguistic feature extraction using independent component analysis. In *Proceedings of IJCNN 2004, International Joint Conference on*

Table 10.2: Most prominent words for three example features (columns) that list clearly related words in both languages.

saksan	values	eroja
ranskan	rauhan	different
germany	demokratian	difference
france	vapauden	välillä
french	democracy	erilaista
german	ihmisoikeuksien	differences
sweden	arvoja	erot
netherlands	solidarity	toisiaan
ranska	peace	disparities
belgian	arvojen	eri
ruotsin	kunnioittaminen	erilaiset
saksa	oikeusvaltion	differ
italian	principles	differing
kingdom	continent	eroavat
...	...	...

Neural Networks, Budapest, Hungary, 25–29 Jul 2004, pp. 279–284.

[3] J.J. Väyrynen, T. Honkela, and A. Hyvärinen. Independent component analysis of word contexts and comparison with traditional categories. In: Jarmo M. A. Tanskanen (ed.), *Proceedings of NORSIG 2004, Sixth Nordic Signal Processing Symposium*, Espoo, Finland, 9–11 Jun 2004, pp. 300–303.

[4] J. J. Väyrynen, L. Lindqvist and T. Honkela. Sparse distributed representations for words with thresholded independent component analysis. In *Proceedings of IJCNN 2007, International Joint Conference on Neural Networks*, Orlando, Florida, 12–17 Aug 2007, pp. 1031–1036.

[5] J. J. Väyrynen, T. Honkela and L. Lindqvist. Towards explicit semantic features using independent component analysis. In: M. Sahlgren and O. Knuttson (eds.), *Proceedings of SCAR 2007 Workshop, Semantic Content Acquisition and Representation*, SICS Technical Report T2007-06, Swedish Institute of Computer Science, Stockholm, Sweden, ISSN 1100-3154, Tartu, Estonia, 24 May 2007, pp. 20–27.

[6] J. J. Väyrynen and T. Lindh-Knuutila. Emergence of multilingual representations by independent component analysis using parallel corpora. In *Proceedings of SCAI 2006, Ninth Scandinavian Conference on Artificial Intelligence*, Espoo, Finland, 25–27 Oct 2006, pp. 101–105.

Computational Cognitive Systems

Chapter 11

Emergence of linguistic and cognitive representations

Timo Honkela, Krista Lagus, Tiina Lindh-Knuutila, Matti Pöllä, Juha Raitio, Sami Virpioja, Jaakko J. Väyrynen and Paul Wagner

11.1 Introduction

Computational Cognitive Systems group conducts research on artificial systems that combine perception, action, reasoning, learning and communication. This area of research draws upon biological, cognitive and social system approaches to understanding cognition. Cognitive systems research is multidisciplinary and benefits from sharing and leveraging expertise and resources between disciplines. For example, statistical machine learning, pattern recognition and signal processing are central tools within computational cognitive systems research. Our research focuses on modeling and applying methods of unsupervised and reinforcement learning.

The general aim is to provide a methodological framework for theories of conceptual development, symbol grounding, communication among autonomous agents, agent modeling, and constructive learning. We have also worked in close collaboration with other groups in our laboratory, for instance, related to multimodal interfaces.

In the following chapters of this report, we describe the main results gained during 2006-07 in the four main thematic areas of the computational cognitive systems research group: *emergence of linguistic and cognitive representations*, *learning social interactions between agents*, *learning to translate* and *knowledge translation and innovation using adaptive informatics*. Each of these areas is described in a section of its own except for emergence of linguistic and cognitive representations that is a shared research area with Multimodal Interfaces research group and the results are mainly described in the chapter on Natural Language Processing. In the following, this area is discussed briefly. In general, the group has benefited strongly from the closeness and the collaboration with other research groups in the center. A notable example of such collaboration is the development of a *speech-to-speech machine translation system protytype*, developed in collaboration with the Multimodal Interfaces research group.

11.2 Research on emergence

The research on emergent linguistic and cognitive representations enables computers to deal with semantics: to process data having certain access to its meaning within multimodal contexts. Our focus is on the analysis and generation of conceptual structures and complex meanings.

The emergence of representations can be considered to consist of the following interrelated tasks: the discovery of elements of representation (e.g. words, morphemes, phonemes), their meaning relations (syntax and semantics), and structures or "rules" of their use in natural utterances (syntax and pragmatics).

An example of work on the first topic is the study of unsupervised machine learning techniques for finding the optimal segmentation of words into sub-word units called morphs, with the intent of finding realizations of morphemes (see Section 10.1 for details). Another example of the use of independent component analysis in discovering meaningful syntactic and semantic features for words (see Section 10.1 for details).

We have studied the emergence of linguistic representations through the analysis of words in contexts using the Independent Component Analysis (ICA). The ICA learns features automatically in an unsupervised manner. Several features for a word may exist, and the ICA gives the explicit values of each feature for each word. In our experiments, we have shown that the features coincide with known syntactic and semantic categories. More detailed description of this research is given in the section on Natural Language Processing in this report.

11.3 Events and projects

An important part of our activity has been the active role in organizing international scientific events. The main activity in 2006-2007 was the organization of the Scandinavian Conference on Artificial Intelligence [1].

In addition to the research areas discussed above, we continue to use the Self-Organizing Map (SOM) where applicable. The most important piece of research based on the SOM during this period was an analysis of conducted for the Academy of Finland. As a continuation to an earlier manually conducted qualitative analysis, the Academy of Finland, one of the country's largest funding agencies, commissioned a study to investigate whether text mining based on the Self-Organizing Map could be used to support assessment of the applications. A collection of 3224 applications was analyzed [2]. A collection of 1331 term candidates was extracted automatically. The 3224 application documents were encoded as term distribution patterns. The SOM algorithm organized the documents into a map in which similar applications are close to each other and in which thematic areas emerged.

Many parts of the research in this new group's agenda have been started during the reported period. Therefore, so far the results have mainly been reported in conferences and as technical reports. However, in early 2008 we were pleased to receive news about four accepted journal papers. Some of the related results are described already in this report, based on publications that have appeared during 2006 and 2007.

Some of the research activities by the group take place in projects funded by Tekes and EU Commission. The most notable examples are the projects Kulta (using adaptive informatics methods to model and simulate changing needs of consumers) and MeIEQ (developing quality labeling methods for medical web resources), discussed in some detail in the section on Knowledge translation and innovation using adaptive informatics.

References

[1] T. Honkela, T. Raiko, J. Kortela and H. Valpola (eds.) (2006). *Proceedings of SCAI'06, the Ninth Scandinavian Conference on Artificial Intelligence*. Finnish Artificial Intelligence Society, 239 p. Espoo.

[2] T. Honkela, M. Klami (2007). Text mining of applications submitted to the Academy of Finland [in Finnish], unpublished report. Helsinki University of Technology, Espoo.

Chapter 12

Learning social interactions between agents

Ville Könönen, Timo Honkela, Tiina Lindh-Knuutila, Mari-Sanna Paukkeri

12.1 Introduction

One important feature of an intelligent agent is its ability to make rational decisions based on its current knowledge of the environment. If the environment of the agent is not static, i.e., there are other active entities, e.g., other agents or humans in the environment, it is crucial to model these entities for making rational decisions.

Our earlier research has been concentrated on the theoretical aspects of the modeling other agents in the reinforcement learning framework. Reinforcement learning is a learning paradigm located between supervised and unsupervised learning. In an environment, the agent takes actions and receives reward signals corresponding to the success of these actions. Correct answers are not directly provided to the agent but it learns features of the environment by continuously interacting with it.

12.2 Applications of multiagent reinforcement learning

Reinforcement learning methods have attained lots of attention in recent years. Although these methods and procedures were earlier considered to be too ambitious and to lack a firm foundation, they have been established as practical methods for solving, e.g., Markov Decision Processes (MDPs). However, the requirement for reinforcement learning methods to work is that the problem domain in which these methods are applied obeys the Markov property. Basically this means that the next state of a process depends only on the current state, not on the history. In many real-world problems this property is not fully satisfied. However, many reinforcement learning methods can still handle these situations relatively well. Especially, in the case of two or more decision makers in the same system the Markov property does not hold and more advanced methods should be used instead. A powerful tool for handling these highly non-Markov domains is the concept of Markov game. In this section we introduce two applications of multi-agent reinforcement learning, namely a dynamic pricing problem and a communication game between agents.

Dynamic pricing

A dynamic pricing scenario is a problem domain that requires planning and therefore it is an ideal testbed for reinforcement learning methods. In the problem, there are two competing brokers that sell identical products to customers and compete on the basis of price. We have modeled the problem as a Markov game and solve it by using two different learning methods. The first method utilizes modified gradient descent in the parameter space of the value function approximator and the second method uses a direct gradient of the parameterized policy function. [1]

Communication game between agents

As another problem, we consider a multiagent system, where unsupervised learning is used in the formation of agents' conceptual models. An intelligent agent usually has a goal. For studying agent based systems formally, it is useful that the goal can be expressed mathematically. Traditional approach is to define a utility function for the agent, i.e. there is a scalar value connected to each possible action measuring the fitness of the action choice for satisfying the goal of the agent. The utility function is often initially unknown and must be learned by interacting with the environment, e.g. by communicating with other agents. [2]

The agents communicate and learn through communication leading into intersubjective sharing of concepts. Communication can be modeled as a mathematical game and the structure of the game can be learnt by using reinforcement learning methods.

References

[1] V. Könönen (2006). Dynamic Pricing Based on Asymmetric Multiagent Reinforcement Learning. *International Journal of Intelligent Systems*, vol. 21, pp. 73–98.

[2] T. Honkela, V. Könönen, T. Lindh-Knuutila and M. Paukkeri (2006). Simulating processes of language emergence, communication and agent modeling. In *Proceedings of STeP-2006, the 12th Finnish Artificial Intelligence Conference*, pp. 129–132. Espoo, Finland.

12.3 Meaning negotiation using simulated naming games

The acquisition of concepts can be viewed as the process of grounding them in language use. How do concepts form, to begin with, when no language exists in the environment? The relationship between cognitive and linguistic development can be pinpointed to the question: how do conceptual emergence and the formation of names for the concepts connect? An associated important question is how an agreement on the use of words is reached in a community of agents. This process of converging towards a shared use of words is called meaning negotiation.

In [4], we studied the emergence of associations between concepts and words with help of a computer simulation and consider the hypothesis that concepts are modeled as areas in some conceptual space (see [1]). Furthermore, we utilized the self-organizing map [3] as a model for an agent's conceptual memory [2]. The feasibility of this conceptual memory model was then studied using multi-agent language games, more specifically, naming games, in which agents learn to associate words to meanings in a communicative setting. Prior to learning word-meaning-associations, a conceptual representation is also learned individually by each agent based on sensory data.

Examples of conceptual maps from simulations with six agents are shown in Fig. 12.1. They show the conceptual space of one agent after the simulation run. The self-organizing map is labeled with the words that have been used during the simulation.

Figure 12.1: An example conceptual map from the simulation. The shades of gray denote distances in the original space: The larger the distance, the darker the color. One can see that there are eight clusters on each map which corresponds to the number of prototypical colors used in the input.

The results of the experiments show clearly that when using the observational game model and the SOM-based conceptual maps (1) the agents learned to communicate successfully on the topics of the games, and (2) a shared lexicon was developed during the simulations. According our definition of successful communication, the agents are also able to communicate successfully and develop a shared lexicon based on adaptation.

References

[1] P. Gärdenfors, *Conceptual Spaces*. MIT Press, 2000.

[2] T. Honkela. Neural nets that discuss: a general model of communication based on self-organizing maps. In S. Gielen and B. Kappen, editors, *Proceedings of ICANN'93, International Conference on Artificial Neural Networks*, pages 408–411, Amsterdam, the Netherlands, September 1993. Springer-Verlag, London.

[3] T. Kohonen (2001). *Self-Organizing Maps*. Third, extended edition. Springer.

[4] T. Lindh-Knuutila, T. Honkela, and K. Lagus (2006). Simulating Meaning Negotiation Using Observational Language Games. In *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication*, pp. 168–179. Springer.

Chapter 13

Learning to translate

Timo Honkela, Mathias Creutz, Tiina Lindh-Knuutila, Sami Virpioja, Jaakko J. Väyrynen

13.1 Introduction

Learning to translate research focuses on developing methods and tools facilitating translations between different languages and even between different dialects or domains. Underpinning development of learning to translate methodology is the fact that contextual, experiential and/or disciplinary diversity impede interpersonal communication and understanding. Our research focuses on the use of unsupervised statistical machine learning. However, in comparison with the traditional approach in statistical machine translation (SMT), we want to take into account known linguistic levels and theories. This does not take place by encoding linguistic knowledge manually to the systems but through architectural choices. For instance, the basic statistical machine translation approach does not properly take into account the morphological or the semantic level. These issues are discussed in the following.

We have applied a method of unsupervised morphology learning to a state-of-the-art phrase-based SMT system [2]. In SMT, words are traditionally used as the smallest units of translation. Such a system generalizes poorly to word forms that do not occur in the training data. In particular, this is problematic for languages that are highly compounding, highly inflecting, or both. An alternative way is to use sub-word units, such as morphemes. We have used the Morfessor algorithm to find statistical morpheme-like units (called morphs) that can be used to reduce the size of the lexicon and improve the ability to generalize. This approach is described more in detail in Section 13.3.

The more general the domain or complex the style of the text the more difficult it is to reach high quality translation. The same applies to natural language understanding. All systems need to deal with problems that relate to the lack of semantic coverage and understanding of the pragmatic level of language. Statistical machine translation systems typically rely on applying Bayes' rule:

> We assign to every pair of strings, s (source) and t (target), in two languages a number $P(t|s)$, which is the probability that a translator, when presented with s, will produce t as the translation. Using Bayes' theorem, one can write $P(s|t) = P(t|s) * P(s)/P(t)$

Thus, in the basic SMT approach, the inputs and outputs are handled only as strings of symbols (consider, e.g., [1, 3]. The system does not receive or deal with information on the meaning of the expressions. To overcome this limitation, there are a number of systems with a hybrid approach, using, for instance, a parser that annotates the training samples with (syntactic and) semantic labels. However, as we wish to minimize the manual effort in development machine translation systems, we have chosen not to use traditional parsers or labeling schemes. Rather, we build on distributional information. Namely, the finding that word co-occurrence statistics, as extracted from text corpora, can provide a natural basis for semantic representations has been gaining growing attention. Words with similar distributional properties often have similar semantic properties. Therefore, it is possible to dynamically build semantic representations of the lexical space through the statistical analysis of the contexts in which words co-occur. In addition to distributional information on the word occurrences in text corpora, also other kinds of contextual information may be used.

The early work by Ritter and Kohonen with artificially generated short sentences as well as contextual information showed the feasibility of the approach outlined above [4]. This work was extended to natural data in [2]. In Section 13.4, we describe how the use of the Self-Organizing Map can be extended to multilingual processing in order to find

semantic grounding for expressions in multiple languages. Some preliminary results on *visual grounding of meaning* are also discussed.

Before addressing the use of unsupervised learning in finding morphological and semantic models that are useful for machine translation, we consider in the following the structural complexity of a number of European languages in Section 13.2. The basic motivation for this analysis lies in the hypothesis that the translation between such two languages is relatively easier that encode information in a similar manner with respect to morphology and syntax. The results of the analysis should thus help in designing translation strategies for automated solutions for various pairs of languages.

References

[1] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, vol. 19(2), pp. 263–311.

[2] T. Honkela, V. Pulkki, and T. Kohonen (1995). Contextual relations of words in Grimm tales, analyzed by self-organizing map. In *Proceedings of ICANN'95, International Conference on Artificial Neural Networks*, vol. II, pp. 3–7. Nanterre, France: EC2.

[3] P. Koehn, F. J. Och, and D. Marcu (2003). Statistical phrase-based translation. In *Proceedings of NAACL'03, North American Chapter of the Association for Computational Linguistics on Human Language Technology*. pp. 48–54. Morristown, NJ, USA: Association for Computational Linguistics.

[4] H. Ritter and T. Kohonen, (1989). Self-organizing semantic maps. *Biological Cybernetics*, vol. 61, no. 4, pp. 241–254.

13.2 Analyzing structural complexity of languages

The European Union has 21 official languages (including Irish from 1st of January 2007), which have approximately 407 million speakers. We have analyzed parallel corpora in these 21 languages using statistical, unsupervised learning methods to study the similarities and differences of the languages in different levels. We have compared these results with traditional linguistic categorizations like division into language groups, morphological complexity and syntactic complexity [3]. The aim of the study has been to evaluate the possibility of using statistical methods in different tasks related to statistical machine translation. For instance, for some language pairs the issues related to morphological analysis may be particularly relevant. For some other language pairs, one may have to pay particular attention to the word order. These kinds of questions can be taken into account when the statistical models to be used are chosen.

Use of compression as a measure for complexity is based on the concept of Kolmogorov complexity. Informally, for any sequence of symbols, the Kolmogorov complexity of the sequence is the length of the shortest algorithm that will exactly generate the sequence and then stop. In other words, the more predictable the sequence, the shorter the algorithm needed is and thus the Kolmogorov complexity of the sequence is also lower [4]. Kolmogorov complexity is uncomputable, but file compression programs can be used to estimate the Kolmogorov complexity of a given file. A decompression program and a compressed file can be used to (re)generate the original string. A more complex string (in the sense of Kolmogorov complexity) will be less compressible. Estimations of complexity using compression has been used for different purposes in many areas. Juola [2] introduces comparison of complexity between languages on morphological level for linguistic purposes.

To get a meaningful interpretation for the order of languages in the word order complexity counting, linguistic literature was consulted for independent figures. Bakker [1] has analyzed flexibility of language's word order, which is based on 10 factors, such as order of verb and object in the language, order of adjective and its head noun, order of genitive and its head noun, etc. The flexibility of the language in Bakker's counting can be given with a numeric value from 0 to 1: if the flexibility figure is close to zero, the language is more inflexible in its word order, if the figure is closer to one, the language is more flexible in its word order. In the information theoretic framework of the compression approach flexibility and inflexibility can be interpreted naturally as higher and lower degrees of complexity, i.e. predictability. In the table below, figures based on Bakker's counting of the flexibility values for the individual languages are given together with values given by compression analysis.

If one compares the figures given by Bakker in column 3 to figures given by compression based calculation in column 6, we can see, that the overall order of the languages based on these independent calculations converge well. The lower end of the scale is quite analogous in both analyses consisting of five same languages with differences in the order. There are also some differences in the orders given by the two analyses. The syntactic complexity of Lithuanian seems to be estimated higher by compression than by Bakker's flexibility value (rank 16 vs. 8). Slovene has also a higher flexibility value than its complexity value (rank 14 vs. 7). Greek is also higher in Bakker's counting than in complexity analysis (rank 17 vs. 11). In our compression calculations Finnish and Estonian are estimated almost equally complex, but in Bakker's analysis Estonian is less complex than Finnish (rank 18 vs. 13).[3]

Bakker's results			Compression results		
1.	fr	0.10	1.	fr	0.66
2.	ga	0.20	2.	es	0.68
3.	es	0.30	3.	pt	0.68
4.	pt	0.30	4.	ga	0.69
5.	it	0.30	5.	it	0.69
6.	da	0.30	6.	en	0.69
7.	mt	0.30	7.	sl	0.71
8.	lt	0.30	8.	nl	0.71
9.	en	0.40	9.	mt	0.72
10.	nl	0.40	10.	da	0.72
11.	de	0.40	11.	el	0.73
12.	sv	0.40	12.	sv	0.75
13.	et	0.40	13.	lv	0.75
14.	sl	0.50	14.	de	0.75
15.	lv	0.50	15.	pl	0.76
16.	sk	0.50	16.	lt	0.76
17.	el	0.60	17.	sk	0.77
18.	pl	0.60	18.	et	0.78
19.	fi	0.60	19.	fi	0.79

References

[1] D. Bakker (1998). Flexibility and Consistency in Word Order Patterns in the Languages of Europe. In Siewierska, A. (ed.): *Constituent Order in the Languages of Europe. Empirical Approaches to Language Typology.* pp. 381–419. Mouton de Gruyter, Berlin New York.

[2] P. Juola (1998) Measuring Linguistic Complexity: the Morphological Tier. *Journal of Quantitative Linguistics*, vol. 5, pp. 206–213.

[3] K. Kettunen, M. Sadeniemi, T. Lindh-Knuutila and T. Honkela (2006). Analysis of EU Languages Through Text Compression. In *Proceedings of FinTAL, the 5th International Conference on NLP*, pp. 99–109. Turku, Finland, August 23–25.

[4] M. Li, X. Chen, X. Li, B. Ma, P. M. B. Vitányi (2004). The Similarity Metric. *IEEE Transactions on Information Theory*, vol. 50, pp. 3250–3264.

13.3 Morphology-Aware Statistical Machine Translation

Statistical machine translation was applied to the direct translation between eleven European languages, all those present in the Europarl corpus, by [1]. An impressive number of 110 different translation systems were created, one for each language pair. Koehn discovered that the most difficult language to translate from or to is Finnish. Finnish is a non-Indo-European language and is well known for its extremely rich morphology. As verbs and nouns can, in theory, have hundreds and even thousands of word forms, data sparsity and out-of-vocabulary words present a huge problem even when large corpora are available.

It appears that especially translating into a morphologically rich language poses an even more substantial problem than translating from such a language. The study also showed that English, which has almost exclusively been used as the target language, was the easiest language to translate into. Thus it is natural to suspect that English as a target language has biased SMT research.

In the following, we describe how we have used morphological information found in an unsupervised manner in SMT [2]. We have tested the approach with the three Nordic languages, i.e., Finnish, Danish and Swedish. Danish and Swedish are closely related languages but differ considerably from Finnish. Danish and Swedish are grammatically very close and much of the vocabulary is shared except for some differences in pronunciation and orthography.

The parallel Europarl corpus [1] of European Parliament Proceedings was used to train our models. Word segmentation models for both source and target languages were trained using Morfessor. At this point, two data set were created for each alignment pair: one with the original word tokens and the other with morph tokens. This allowed us to create a comparable baseline system. We used standard state-of-the-art n-gram language models trained with the target language text. Phrase-based translation models were trained with Moses, an open-source statistical machine translation toolkit [3]. A phrase-based system translates short segments of consecutive words in contrast to word-based translation, which translates one word at a time. Phrases enable more natural language generation and flexibility in translating, for instance, idioms, collocations, inflected words forms and compound words in which the number of words may not stay the same across translation. The parameters of word-based and morph-based system were the same, except for the maximum number of tokens in a phrase, that was higher with morph-based systems to cover approximately the same number of words than word-based systems.

Figure 13.1 shows an example how in addition to the the different tokens, morph-based translation first segments words into morphs and finally after the translation constructs words from the translated morphs. Having morph tokens lowers the type counts greatly compared to words. The segmentation naturally increases tokens counts slightly. Reduced type counts help with sparse data, and this was especially prominent with Finnish. On the other hand, increased token counts seem to make the word alignment and translation process more complicated.

In the word-based translation model, only the words that were present in the training data can be translated. The other words are left untranslated, even though they may simply by an inflected form of a known word. Thus we expected to get less untranslated words with the morph-based system. This was true, as shown in Table. 13.1. An examination of the untranslated words reveals that a higher number of compound words and inflected word forms are left untranslated by the word-based systems.

As in most of the recent studies, we have used the BLEU scores [4] for quantitative evaluation. BLEU is based on the co-occurrence of n-grams between a produced transla-

a	flera reglerande åtgärder behöver införas .			
b	flera	reglerande åtgärder	behöver införas	.
c	eräitä	sääntelytoimia	on toteutettava	.
d	eräitä sääntelytoimia on toteutettava .			

e	flera reglerande åtgärder behöver införas .				
f	flera_0 reglera^*_0 nde_+ åtgärd^*_0 er_+ behöv^*_0 er_+ in^*_- föra^*_0 s_+ $\cdot_0$				
g	flera_0	reglera^*_0 nde_+	åtgärd^*_0 er_+	behöv^*_0 er_+ in^*_- föra^*_0 s_+	$\cdot_0$
h	erä^*_0 itä_+	sääntely^*_0	toimi^*_0 a_+	on_0 toteute^*_0 tta^*_+ va_+	$\cdot_0$
i	erä^*_0 itä_+ sääntely^*_0 toimi^*_0 a_+ on_0 toteute^*_0 tta^*_+ va_+ $\cdot_0$				
j	eräitä sääntelytoimia on toteutettava .				

Figure 13.1: Examples of word-based and morph-based Finnish translations for the Swedish sentence "Flera reglerande åtgärder behöver införas ." (*Several regulations need to be implemented .*) The top figure shows the word-based translation process with the source sentence (a), the phrases used (b) and their corresponding translations (c), as well as the final hypothesis (d). The bottom figure illustrates the morph-based translation process with the source sentence as words (e) and as morphs (f), the morph phrases used (g) and their corresponding translations (h), as well as the final hypothesis with morphs (i) and words (j). Each morph if either a prefix ($-$), a stem (0) or a suffix ($+$), marked by the lowerscript. A superscript ($*$) marks the morphs that are not the last one in the word.

word / morph	→ Danish	→ Finnish	→ Swedish
Danish →		128 / 31	74 / 12
Finnish →	189 / 41		195 / 44
Swedish →	76 / 21	132 / 42	

Table 13.1: Number of sentences with untranslated words out of 1 000 with word-based and morph-based phrases.

tion and a reference translation. BLEU score has been critized, for instance, as in some cases human evaluation gives grossly different results. It is also clear that for morphologically rich languages, such as Finnish, it is harder to get good scores on a word-token based evaluation method. In Table 13.2, the differences between the scores for word-based and morph-based systems are shown, with statistically significant differences highlighted. According to these results, the translations based on morph phrases were slightly worse, but only in two cases the decrease was statistically significant.

References

[1] P. Koehn (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X, 10th Machine Translation Summit*, pp. 79–86. Phuket, Thailand, Sep 13–15.

[2] S. Virpioja, J. J. Väyrynen, M. Creutz and M. Sadeniemi (2007). Morphology-Aware Statistical Machine Translation Based on Morphs Induced in an Unsupervised Manner. In *Proceedings of MT Summit XI, 11th Machine Translation Summit*, pp. 491–498. Copenhagen, Denmark, Sep 10–14.

	→ Danish	→ Finnish	→ Swedish
Danish →		-0.60	-0.52
Finnish →	-1.23		**-2.14**
Swedish →	-0.46	**-1.14**	

Table 13.2: Absolute changes in BLEU scores from word-based translations to morph-based translations. The maximum phrase length was 7 for words and 10 for morphs. 4-gram language models were used for both. Statistically significant differences are marked with boldface fonts.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, & E. Herbst (2007). Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of ACL, demonstration session*. Czech Republic, June.

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pp. 311–318. Morristown, NJ, USA.

13.4 Self-Organizing Semantic Representations for Machine Translation

Discussing the fundamental problems of translation, Quine has presented a situation in which one is confronted with a situation in which one must attempt to make sense of the utterances and gestures that the members of a previously unknown tribe make [3]. Quine claimed that it is impossible, in such a situation, to be absolutely certain of the meaning that a speaker of the tribe's language attaches to an utterance. For example, if a speaker sees a rabbit and says "gavagai", is she referring to the whole rabbit, to a specific part of the rabbit, or to a temporal aspect related to the rabbit. Even further, if one considers the symbol grounding problem [1], there can practically even be an infinite number of conceptualizations of the situation. Maybe the members of the tribe not only consider the whole rabbit or some parts or aspects of it as potentially relevant points of reference but, e.g., due to their cultural context they consider some other patterns of perception. Namely, considering the complex pattern recognition process, it is far from trivial to create a perception of a rabbit from the raw visual and auditory input.

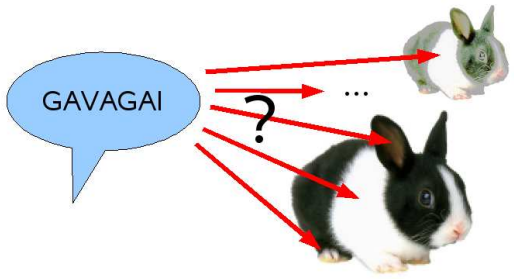


Figure 13.2: An illustration of the reference problem (see text for details).

Quine mentions that one can form manuals of translation [3]. The observer examines the utterances as parts of the overall linguistic behavior of the individual, and then uses these observations to interpret the meaning of all other utterances. Quine continues that there will be many such manuals of translation since the reference relationship is indeterminate. He allows that simplicity considerations not only can be used to choose between competing manuals of translation but that there is even a remote possibility of getting rid of all but one manual.

It seems that propositional logic as the underlying epistemological framework unnecessarily complicates the consideration. For Quine it was necessary to consider a number of logically distinct manual of translation hypotheses. However, if one considers the issue within the framework of statistics, probability theory and continuous multidimensional representations of knowledge, one can consider the conditional probability of different hypotheses and partial solutions which do not need to be logically coherent. Moreover, the search for translation mappings can be seen as a process that may (or may not) converge over time. For Quine meaning is not something that is associated with a single word or sentence, but is rather something that can only be attributed to a whole language. The resulting view is called semantic holism. In a similar fashion, the self-organizing map specifies a holistic conceptual space. The meaning of a word is not based on some definition but is the emergent result of a number of encounters in which a word is perceived or used in some context. Moreover, the emergent prototypes on the map are not isolated instances but they influence each other in the adaptive formation process.

Finding a mapping between vocabularies of two different languages, the results of a

new experiment are reported in the following. Maps of words are often constructed using distributional information of the words as input data. The result is that the more similar the contexts in which two words appear in the text, the closer the words tend to be on the map. We have extended this basic idea to cover the notion of context in general. We have considered the use of a collection of words in two languages, English and German, in a number of contexts. In this experiment, the contexts were real-life situations rather than some textual contexts.

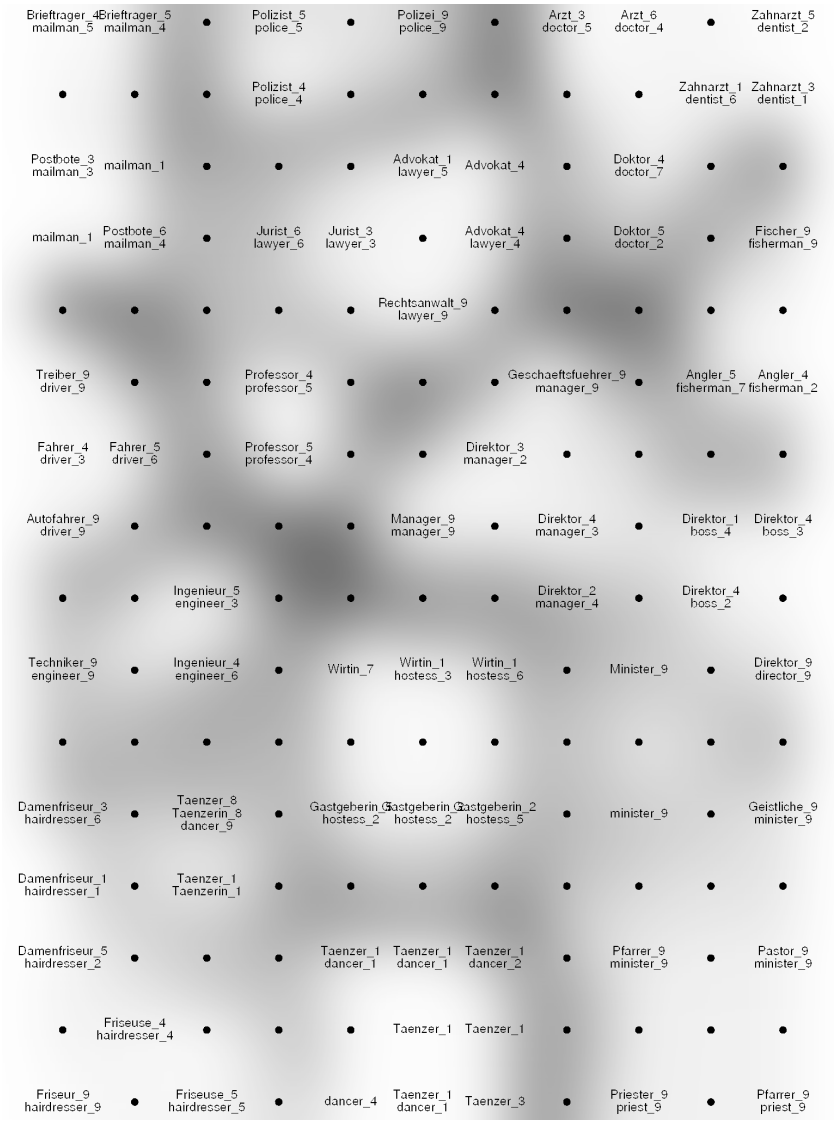


Figure 13.3: An illustration of the reference problem (see text for details).

Figure 13.3 presents the order of some words on a self-organizing map that serves simultaneously two purposes. First, it has organized different contexts to create a semantic landscape. Second, the map includes a mapping between the English and German words used in the analysis. The input for the map consists of words and their contexts. The German vocabulary includes 33 words (Advokat, Angler, Arzt, Autofahrer, ..., Zahnarzt) and the English vocabulary of 17 words (boss, dancer, dentist, director, ..., professor). For each word, there is a assessment by 10 to 27 subjects indicating the degree of suitability for the word to be used in a particular context. The number of contexts used was 19.

The map shows that those words in the two languages that have similar meaning are close to each on the map. In this particular experiment, the German subjects were usually using a larger vocabulary. Therefore, in many areas of the map, a particular conceptual area is covered by one English word (for instance, "doctor" or "hairdresser") and by two or more German words (for instance, "Arzt" and "Doktor" or "Friseur", "Friseuse" and "Damenfriseur").

The research on content-based information retrieval and analysis (see Chapter 7) provides a solid basis for future research in "translation through images". Some initial experiments show that names of concrete objects in different languages can be mapped with each other without any intermediate linguistic/symbolic representation (see [5] for details). In general, these results support the idea of symbol grounding [4].

References

[1] S. Harnad (1990). The symbol grounding problem. *Physica D*, vol. 42, pp. 335–346.

[2] T. Honkela (2007). Philosophical Aspects of Neural, Probabilistic and Fuzzy Modeling of Language Use and Translation In *Proceedings of IJCNN 2007, International Joint Conference on Neural Networks*. Orlando, Florida, Aug 12–17.

[3] W. Quine, (1960). *Word and Object*. MIT Press.

[4] M. Sjöberg, J. Laaksonen, M. Pöllä, T. Honkela (2006). Retrieval of Multimedia Objects by Combining Semantic Information from Visual and Textual Descriptors. In *Proceedings of ICANN 2006, International Conference on Artificial Neural Networks*, pp. 75–83. Athens, Greece.

[5] M. Sjöberg, V. Viitaniemi, J. Laaksonen, T. Honkela (2006). Analysis of Semantic Information Available in an Image Collection Augmented with Auxiliary Data. In *Proceedings of IFIP, Conference on Artificial Intelligence Applications and Innovations*, pp. 600–608. Athens, Greece.

Chapter 14

Knowledge translation and innovation using adaptive informatics

Timo Honkela, Mikaela Klami, Matti Pöllä, Ilari Nieminen

14.1 Introduction

Knowledge translation can be defined as the process of supporting the uptake of research in a manner that improves the practices in the society and in industries through improved understanding, processes, services, products or systems. The term knowledge translation is used rather widely in health care: knowledge translation activities include: (1) research into the mechanisms of knowledge translation, and (2) evidence-based translation of knowledge (e.g., knowledge dissemination, technology transfer, knowledge management, knowledge utilization, synthesis of research results within a local or global context, development and application of consensus guidelines). Knowledge translation in any discipline requires a reciprocal relationship between research and practice. The goals of knowledge translation are to enhance competitiveness, innovation and quality of services and products. Adaptive informatics can provide tools for supporting knowledge translation and innovation.

14.2 Statistical machine learning systems as traveling computational models

In collaboration with the department of philosophy at University of Helsinki, we have considered statistical machine learning models from a more general level focusing on the question on what can be said about them as scientific methods and tools. This question has strategic interest when the use of these model is considered. Within science and technology studies, Dr. Tarja Knuuttila has for some time conducted research on studying scientific models as epistemic artifacts. Earlier her interest has focused on explicit models in computational linguistics such as parsers. Traditional linguistic models can be considered to be first-order models of language. Unsupervised learning methods, on the other hand, can be called second-order models: They do not model the phenomenon directly but through a process of emergence. In the following, the collaboration work is described more in detail based on [4] (see also [3]). We focus on neural network models and specifically on the Self-Organizing Map (SOM) [2]. The discussion, however, applies basically to any unsupervised statistical machine learning method.

Traditionally, it has been thought that models are primarily models of some target systems, since they represent partially or completely the target systems. Sometimes computational models not only have various roles within a scientific domain, but also travel across scientific disciplines. A travelling computational template is a computational method that has a variety of applications in different scientific domains [1]. Neural networks are good examples of such traveling templates. Initially, most of them were inspired by the idea of looking at brains as a model of a parallel computational device, but nowadays neural networks are applied in several different scientific domains, not all of which belong to the domain or neuroscience.

That a model can have so many and various applications, raises opinion some significant philosophical issues concerning the nature of models and how they give us knowledge. Both questions have been answered by philosophers of science by reverting to representation. On one hand models are considered to be representations, on the other hand they are thought to give us knowledge because they represent.

In the recent research in the philosophy of science it has been shown that neither isomorphism nor similarity can provide an adequate analysis of scientific representation. They lead to well-known problems. Firstly, the isomorphism view in fact assumes that there is no such thing as false representation, either the model and its target system are isomorphic, or then they are not, in which case there is no representation either. Secondly, both isomorphism and similarity are symmetric relations, which runs counter our intuitions about representation: we want a model to represent its target system but not vice versa. Both problems appear to be solved once the pragmatic aspects of representation are taken into account. The users' intentions create the directionality needed to establish a representative relationship; something is being used and/or interpreted as a model of something else. Taking into account human agency introduces also indeterminateness into the representative relationship: human beings as representers are fallible. Consequently, pragmatic approaches to representation solve many problems of the structuralist notion of representation-but this comes with a price. When representation is grounded primarily on the specific goals and representing activity of humans as opposed to the respective properties of the representative vehicle and its target system, nothing very substantive can be said about representation in general: There is nothing in the nature of the representation (the model) and its target system that guarantees the representational relationship between the two.

More importantly, even though the SOM were originally inspired by the neurophys-

iological structures of the cortex that does not explain their success in other domains. What is more, when SOMs are used in the fields of inquiry that lie quite afar from the neurophysiological research of the cortex, to conceive SOMs as representations becomes often vague. Instead, the SOM models reveal statistical structure in the data. To do this they rely on a "neurally inspired" algorithm, but this fact does not really make the SOM a representation of a neural representation of the domain of interest. If it represents anything, then it represents the data in a certain way. In this SOM models are alike simulation models on general, since often they are first and foremost appreciated for the output representations they produce. There was also a specific reason to consider the SOM as a sample of a neural network model rather than for instance, a backpropagation algorithm. Namely, as the SOM applies unsupervised learning paradigm, the end result of the analysis reflects relatively more the contents of the data than the supervised learning model sthat impose predetermined output categories on the analysis.

They should rather be conceptualized as multifunctional epistemic artifacts. More generally, the traditional philosophical view according to which models are first and foremost representations of some pre-defined target systems does not capture what seems to us the characteristic feature of modeling: the use of inherently cross-disciplinary computational templates.

References

[1] P. Humphreys. (2002). Computational Models. *Philosophy of Science*, vol. 69, pp. S1-S11.

[2] T. Kohonen. (2001). *Self-Organizing Maps*. 3rd edition, Springer.

[3] T. Knuuttila and T. Honkela. (2005). Questioning External and Internal Representation: The Case of Scientific Models. In L.Magnani (ed.) *Computing, Philosophy, and Cognition*, pp. 209–226. London: King's College Publishing.

[4] T. Knuuttila, A.-M. Rusanen and T. Honkela. (2007). Self-Organizing Maps as Travelling Computational Templates. In *Proceedings of IJCNN 2007, International Joint Conference on Neural Networks*, pp. 1231–1236. Orlando, Florida, Aug 12–17.

14.3 Modeling and simulating practices

In collaboration with Prof. Mika Pantzar (Helsinki School of Economics, on leave from National Consumer Research Centre), we have been developing a simulation model called *Pracsim*[3]. The systems demonstrates the basic concepts of practice theory. The theory is developed by Pantzar in collaboration with Prof. Elizabeth Shove (Lancaster University). In the theory, it is assumed that practices consist of three basic elements: material (materials, technologies and tangible, physical entities), image (domain of symbols and meanings), and skill (competence, know-how and techniques) [1]. Practices come into existence, persist and disappear when links between these foundational elements are made, sustained or broken: material, image and skill co-evolve. For instance, in the case of Nordic walking, Walking sticks are integrated to produce a proper Nordic Walking technique (linking material objects with skills). Furthermore, images of safety, fitness and nature can be integrated into the sticks themselves (linking image and material object) [2]. The basic motivation in considering practices as an application domain for adaptive informatics methods raises from its richness and complexity. The dynamics and conceptual content related to the phenomena of everyday practices sets a clear challenge for methodology development.

Theories on human action are often constructed either in such a way that the emphasis is on the social or on the individual level. Practice theory aims to build a bridge between these points of view. Based on the theory, Pracsim simulation system consists of two main parts: Simulation of practice dynamics and simulation of associated human population, the members of which adopt practices based on a variety of principles. Pracsim is an example of social simulation that refers to a general class of strategies for understanding social dynamics using computers to simulate social systems.

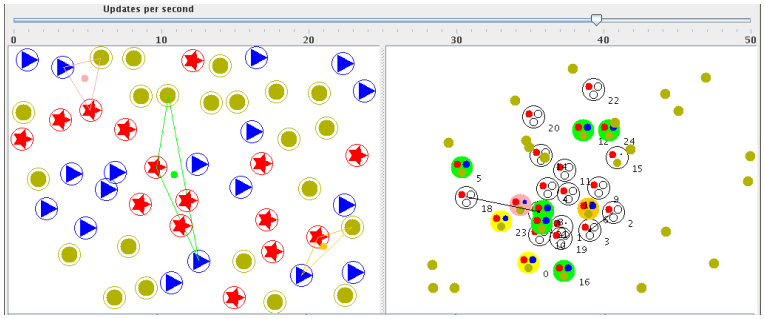


Figure 14.1: A screenshot of the Pracsim simulation. The symbols on the left hand side of the screen refers to the three different elements in the practice theory. The groups of three elements that are linked form a living practice. The right hand side of the screen includes a number of individuals, some of which have adopted a practice. Communication and diffusion within the community of agents is visualized with the links between the individuals. Small items in the "world" are instances of some material.

The pracsim simulation environment consists of a "world" in which a collection of items interact with each other. Following the practice theory, the items are either material, image and skill. Practices can be linked together into systems of practices. These systems are visualized by links between the participating practices. [3]

Pracsim system is applied in a Tekes project called Kulta that develops and applies methods that can be used in understanding, conceptualizing and anticipating the changing needs of consumers. The conceptual models of the practice theory are applied to analyze

changes in the consumer society. These models are applied in the context of developing the business models of different kinds of companies. The data gained in this research and developing process are then analyzed and modeled using adaptive informatics methods. The research is focused on the strategic decision making within companies, and how new modeling techniques can be used in these processes. We have also been studied the usability of the modeling and simulation methods.

References

[1] M. Pantzar and E. Shove. (2008). *The Choreography of Everyday Life: Towards an Integrative Theory of Practice*. Forthcoming.

[2] E. Shove and M. Pantzar. (2005). Consumers, producers and practices: understanding the invention and reinvention of Nordic Walking. *Journal of Consumer Culture*, 1/2005, pp. 43-64.

[3] L. Lindqvist, T. Honkela and M. Pantzar. (2007). Visualizing Practice Theory through a Simulation Model. Helsinki University of Technology, Publications in Computer and Information Science, Report E9.

14.4 Analysis of interdisciplinary Text Corpora

We have presented means for analyzing text documents from various areas of expertise to discover groups of thematically similar texts with no prior information about the topics. These results show how a relatively simple keyword analysis combined with a SOM projection can be very descriptive in terms of analyzing the contextual relationships between documents and their authors.

Our analysis of text documents attempts to extract information about the area of expertise of the document using a set of keywords which are extracted from the documents automatically. To extract relevant keywords for each text document the frequency of each word is examined as an indicator of relevance. As the word frequency distribution for any source of natural language has a Zipf'ian form, special care has to be taken to filter out words, which occur frequently in all documents but are irrelevant in describing the topics. After the keyword extraction phase the documents are analyzed using a SOM projection of the keyword usage of the documents.

We selected two sets of documents from two distinctive fields of expertise: the first corpus A was collected from scientific articles published in the proceedings of the AKRR'05 conference with the topics of the papers mainly in the fields of computer science, cognitive science, language technology and bioinformatics. Corpus B consists of a collection of articles published by the Laboratory of Environmental Protection at Helsinki University of Technology with the topics ranging from social sciences to environmental managing.

Our experiments have shown that a combination of an automatic keyword extraction scheme combined with a clustering algorithm can be effective in describing the mutual similarities of text documents. Our statistical approach to the task has the additional benefit of being independent of the language that is being processed as no prior information about the processed language syntax is encoded into the algorithm.

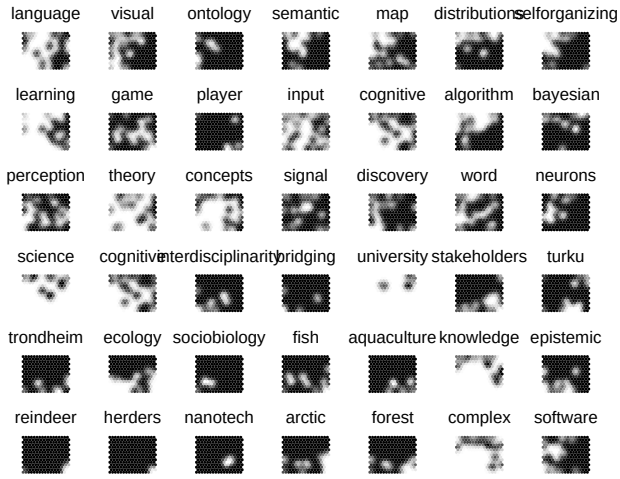


Figure 14.2: Component plane visualization of 42 keywords used in the analysis. Light shade corresponds to the occurrence of the keyword.

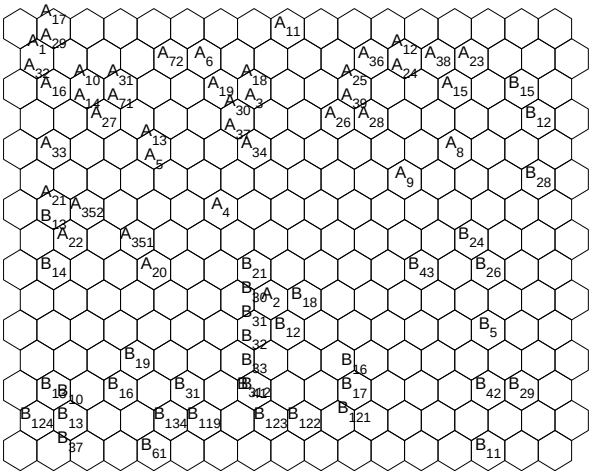


Figure 14.3: The documents of corpora A and B in a SOM projection. The two collections stem from the research of two different research groups which can clearly be seen in the clustering structure.

References

[1] M. Klami and T. Honkela. (2007). Self-Organized Ordering of Terms and Documents in NSF Awards Data. In *Proceedings of WSOM'07, Workshop on Self-Organizing Maps*. Bielefeld, Germany, Sep 3–6.

[2] M. Pöllä, T. Honkela, H. Bruun, and A. Russell. (2006). Analysis of interdisciplinary text corpora. *Proceedings of SCAI'06, the 9th Scandinavian Conference on Artificial Intelligence*, pp. 17–22. Espoo, Finland, Oct 25–27.

14.5 Quality analysis of medical web content

As the number of medical web sites in various languages increases, it is more than necessary to establish specific criteria and control measures that give the consumers some guarantee that the health web sites they are visiting, meet a minimum level of quality standards and that the professionals offering the information on the web site are responsible for its contents and activities.

We are a partner in a EU-funded project called MedIEQ that develops methods and tools for the quality labelling process in medical web sites. MedIEQ will deliver tools that crawl the Web to locate medical web sites in seven different European languages (Spanish, Catalan, German, English, Greek, Czech, and Finnish) in order to verify their content using a set of machine readable quality criteria. MedIEQ tools will monitor already labelled medical sites alerting labelling experts in case the sitesâ content is updated against the quality criteria, thus facilitating the work of medical quality labelling agencies. The overall objective of MedIEQ is to advance current medical quality labelling technology, drawing on past and original research in the area. Our work on automatic keyphrase extraction is used as a component of the MedIEQ AQUA system where relevant terminology about the content of medical web sites is used to facilitate the work of the human expert.

References

[1] T. Honkela, M. Pöllä (2006). Describing Rich Content: Future Directions for the Semantic Web. In *Proceedings of STeP 2006, Finnish Artificial Intelligence Conference*, pp. 143–148, Finnish Artificial Intelligence Society. Espoo, Finland, Oct 26–27.

[2] M. A. Mayer, V. Karkaletsis, K. Stamatakis, A. Leis, D. Villarroel, C. Dagmar, M. Labsky, F. López-Ostenero, T. Honkela (2006). MedIEQ -Quality Labelling of Medical Web Content Using Multilingual Information Extraction. *Studies in Health Technology and Informatics* vol. 121, pp. 183–190.

Adaptive Informatics Applications

Chapter 15

Intelligent data engineering

Olli Simula, Jaakko Hollmén, Kimmo Raivio, Miki Sirola, Timo Similä, Mika Sulkava, Pasi Lehtimäki, Jarkko Tikka, Jukka Parviainen, Jaakko Talonen, Golan Lampi, Mikko Multanen, Tuomas Alhonnoro, Risto Hakala

15.1 Failure management with data analysis

Miki Sirola, Jukka Parviainen, Jaakko Talonen, Golan Lampi, Tuomas Alhonnoro, Risto Hakala, Timo Similä

Early fault detection with data-analysis tools in nuclear power plants is one of the main goals in NoTeS-project (test case 4) in TEKES technology program MASI. The industrial partner in this project is Teollisuuden Voima Oy, Olkiluoto nuclear power plant. Data analysis is carried out with real failure data, training simulator data and design based data, such as data from isolation valve experiments. A control room tool, visualization tools and various visualizations are under development.

A toolbox for data management using PCA (Principal Component Analysis) and WRLS (Weighted Recursive Least Squares) methods has been developed [1]. Visualizations for e.g. trends, transients, and variation index to detect leakages are used. Statistically significant variables of the system are detected and statistical properties and important visualizations are reported. Data mining methods and time series modelling are combined to detect abnormal events.

X-detector tool based on feature subset selection has been developed. The idea is to do real-time monitoring and abnormality detection with efficient subsets. Measuring dependencies and cluster separation methods are used in variable selection in this visualization tool.

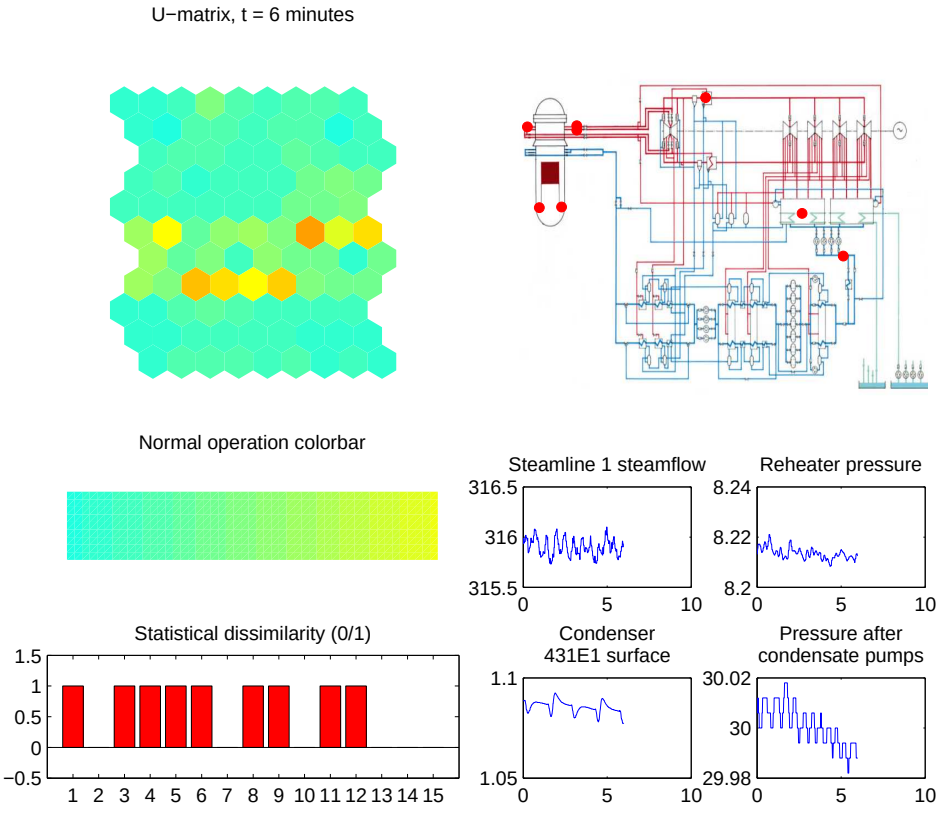


Figure 15.1: X-detector tool user interface: leakage in the main circulation pump. SOM visualization combined with statistical Kolmogorov-Smirnov test, process flow diagram and selected process variable graphs.

Decision support prototype DERSI for failure management in nuclear power plants is under development. It is a control room tool for operator or analysis tool for expert user. It combines neural methods and knowledge-based methods. DERSI utilizes Self-Organizing Map (SOM) method and gives advice by rule-based reasoning. The operator is provided by various informative decision support visualizations, such as SOM maps for normal data and failure data, state U-matrix, quantization error for both component level and state U-matrix, time-series curves and progress visualizations. DERSI tool has been tested in fault detection and separation of simulated data [2].

A separate study of process state and progress visualizations using Self-Organizing Map was also done [3]. All visualizations developed in the project will be collected to make a first proposal for wide monitoring screens.

References

[1] J. Talonen. Fault Detection by Adaptive Process Modeling for Nuclear Power Plant. Master's thesis, Helsinki University of Technology, 2007.

[2] M. Sirola, G. Lampi, and J. Parviainen. Failure detection and separation in SOM based decision support. In *Workshop on Self-Organizing Maps*, Bielefeld, Germany, 2007. WSOM.

[3] R. Hakala, T. Similä, M. Sirola, and J. Parviainen. Process state and progress visualization using self-organizing map. In *International Conference on Intelligent Data Engineering and Automated Learning*, Burgos, Spain, September 2006. IDEAL.

15.2 Cellular network performance analysis

Kimmo Raivio, Mikko Multanen, Pasi Lehtimäki

Structure of mobile networks gets more and more complicated when new network technologies are added to the current ones. Thus, advanced analysis methods are needed to find performance bottlenecks in the network. Adaptive methods can be utilized, for example, to perform hierarchical analysis of the networks, detecting anomalous behavior of network elements and to analyse handover performance in groups of mobile cells.

Combination of the Self-Organizing Map and hierarchical clustering methods can be utilized to split the analysis task into smaller subproblems in which detection and visualization of performance degradations is easier. The method consists of successive selection of a set of cellular network performance indicators and hierarchical clustering of them. Initially only a couple of key performance indicators are utilized and later some more specific counters are used. Thus, the root cause of degradation is easier to find [1]. The method can be utilized both in general network perfromance analysis and in more specific subareas like soft handover success rate [3].

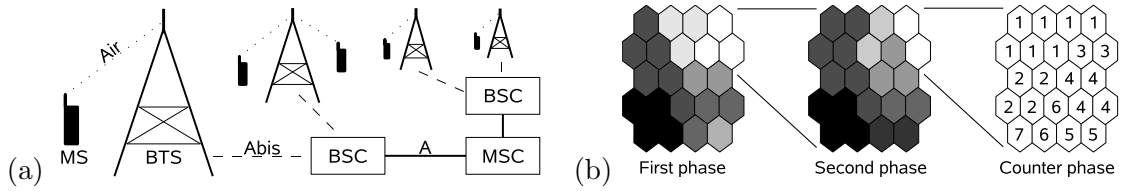


Figure 15.2: Architecture of a cellular network (a) and simple view of the hierarchical analysis algorithm (b).

In outlier detection as well neural as statistical methods can be used to find out network elements with decreased performance or otherwise anomalous traffic profile. Statistical approaches may include both parametric and non-parametric methods. An example of parametric method is Gaussian mixture model. Correspondingly, nearest-neighbor and Parzen windows are non-parametric methods. A neural method called Neural gas is very similar to the statistical approaches and it can be used also in this task [2].

It can be said, that neural and other learning methods can be utilized in the analysis of complicated performance degradation problems in cellular networks. The analysis tools can be built in a way to require only a minimal amount of knowledge of the network itself.

References

[1] M. Multanen, P. Lehtimäki, and K. Raivio. Hierarchical analysis of GSM network performance data. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, pages 449–454, Bruges, Belgium, April 26 - 28 2006.

[2] M. Multanen, K. Raivio, and P. Lehtimäki. Outlier detection in cellular network data exploration. In *Proceedings of the 3rd International Workshop on Performance Analysis and Enhancement of Wireless Networks (PAEWN)*, Okinawa, Japan, March 25 - 28 2008.

[3] K. Raivio. Analysis of soft handover measurements in 3G network. In *Proceedings of the 9th ACM Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 330–337, Torremolinos, Malaga, Spain, October 2 - 6 2006.

15.3 Predictive GSM network optimization

Pasi Lehtimäki, Kimmo Raivio

In this study, the focus is on the final step of the mobile network monitoring procedure, that is, on making adjustments to configuration parameters so that the amount of predictable, regularly occurring performance degradations or faults is minimized. In order to automate the configuration parameter optimization, a computational method to evaluate the performance of alternative configurations must be available. In data-rich environments like cellular networks, such predictive models are most efficiently obtained with the use of past data records.

In blocking prediction, the interest is to compute the number of blocked requests at different conditions. This can be based on the use of well known Erlang-B formula. The expected value for the number of blocked requests is obtained by multiplying the number of arriving requests with the blocking probability, leading to $B = \lambda p(N_c|\lambda, \mu, N_c)$. The expected value for the congestion time is $C = p(N_c|\lambda, \mu, N_c)$ and the expected value for the number of channels in use is $M = \sum_{n=0}^{N_c} np(n|\ \lambda, \mu, N_c)$.

In [2], it was shown that the Erlang-B formula does not provide accurate predictions for blocking in GSM networks if low sampling rate measurements of arrival process are used in the model. More traditional regression methods can be used for the same purpose with the assist of knowledge engineering approach in which Erlang-B formula and regression methods are combined. With the use of Erlang-B formula, the dependencies between B, C and M that remain the same in each base station system need not be estimated from data alone. The data can be used to estimate other relevant and additional parameters that are required in prediction. In [2] and [1], a method to use Erlang-B formula and measurement data to predict blocking is presented. The regression techniques are used to estimate the arrival rate distribution describing the arrival process during short time periods. The Erlang-B formula is used to compute the amount of blocking during the short time periods.

Suppose that the time period is divided into N_s segments of equal length. Also, assume that we have a vector $\boldsymbol{\lambda} = [0 \ \ 1\Delta_\lambda \ \ 2\Delta_\lambda \ \ \ldots \ \ (N_\lambda - 1)\Delta_\lambda]$ of N_λ possible arrival rates per segment with discretization step Δ_λ. Let us denote the number of blocked requests during a segment with arrival rate λ_i with $B_i = \lambda_i p(N_c|\lambda_i, \mu, N_c)$, where $p(N_c|\lambda_i, \mu, N_c)$ is the blocking probability given by the Erlang distribution. Also, the congestion time and the average number of busy channels during a segment with arrival rate λ_i are denoted with $C_i = p(N_c|\lambda_i, \mu, N_c)$ and $M_i = \sum_{n=0}^{N_c} np(n|\lambda_i, \mu, N_c)$. In other words, the segment-wise values for blocked requests, congestion time and average number of busy channels are based on the Erlang-B formula.

Now, assume that the number of segments with arrival rate λ_i is θ_i and $\sum_i \theta_i = N_s$. Then, the cumulative values over one hour for the number of requests T, blocked requests B, congestion time C and average number of busy channels M can be computed with

$$\begin{bmatrix} \lambda_1 & \lambda_2 & \ldots & \lambda_{N_\lambda} \\ B_1 & B_2 & \ldots & B_{N_\lambda} \\ \frac{C_1}{N_s} & \frac{C_2}{N_s} & \ldots & \frac{C_{N_\lambda}}{N_s} \\ \frac{M_1}{N_s} & \frac{M_2}{N_s} & \ldots & \frac{M_{N_\lambda}}{N_s} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{N_\lambda} \end{bmatrix} = \begin{bmatrix} T \\ B \\ C \\ M \end{bmatrix} \tag{15.1}$$

or in matrix notation $\mathbf{X}\boldsymbol{\theta} = \mathbf{Y}$.

Now, the problem is that the vector $\boldsymbol{\theta}$ is unknown and it must be estimated from the data using the observations of $\mathbf{Y}$ and matrix $\mathbf{X}$ which are known a priori. Since the output

vector $\mathbf{Y}$ includes variables that are measured in different scales, it is necessary to include weighting of variables into the cost function. By selecting variable weights according to their variances estimated from the data, the quadratic programming problem

$$\min_{\boldsymbol{\theta}} \quad \left\{ \frac{1}{2} \boldsymbol{\theta}^T \mathbf{H} \boldsymbol{\theta} + \mathbf{f}^T \boldsymbol{\theta} \right\} \tag{15.2}$$

$$w.r.t \quad 0 \leq \theta_i \leq N_s, \ i = 1, 2, ..., N_\lambda, \tag{15.3}$$

$$\sum_{i=1}^{N_\lambda} \theta_i = N_s \tag{15.4}$$

is obtained where $\mathbf{f} = -\mathbf{X}^T \mathbf{W}^T \mathbf{W} \mathbf{Y}$ and $\mathbf{H} = \mathbf{X}^T \mathbf{W}^T \mathbf{W} \mathbf{X}$ include the weighting matrix $\mathbf{W}$. In other words, the goal is to find the vector $\boldsymbol{\theta}$ that provides the smallest prediction errors for variables T, B, C and M.

The optimization problem could be solved for each of the N_d observation vectors separately, leading to N_d solution vectors $\boldsymbol{\theta}$ for hour h. Since we are interested in long-term prediction of blocking, we should somehow combine the solution vectors so that behavior common to all solution vectors are retained and non-regular properties of the demand are given less attention.

Let us denote the ith solution vector for hour h with $\boldsymbol{\theta}_h^{(i)}$ and the jth element of the corresponding solution vector with $\theta_{jh}^{(i)}$. Since $\theta_{jh}^{(i)}$ described the number of segments with arrival rate $\lambda = \lambda_j$ during ith observation vector at hour h, the probability for a random segment during ith observation period to have an arrival rate $\lambda = \lambda_j$ can be computed from $\theta_{jh}^{(i)}$ with $p_{jh}^{(i)} = \theta_{jh}^{(i)} / N_s$, where N_s is the number of segments in a period.

The probability for observing a segment with arrival rate $\lambda = \lambda_j$ at hour h would become

$$p_{jh} = \frac{1}{N_d N_s} \sum_{i=1}^{N_d} \theta_{jh}^{(i)}. \tag{15.5}$$

Now, the arrival rates λ_j and their probabilities p_{jh} for hour h form a probabilistic model. Let us define a column vector

$$\underset{seg \mapsto hour}{\boldsymbol{\theta}_h} = \mathbf{p}_h N_s \tag{15.6}$$

that maps the segment-wise candidate arrival rates λ_j to the total number of arrived requests T in a single one hour time period with

$$T = \boldsymbol{\lambda} \underset{seg \mapsto hour}{\boldsymbol{\theta}_h}. \tag{15.7}$$

Note that the parameter vector $\boldsymbol{\theta}_{h,seg \mapsto hour}$ can also be used to map the vector $\mathbf{B} = [B_1 \ B_2 \ \ldots \ B_{N_\lambda}]$ of segment-wise blocking candidates to the total number of occurrences of blocked requests during one period. Similarly, the cumulative values for the average number of busy channels and the congestion time can be computed.

References

[1] P. Lehtimäki. A model for optimisation of signal level thresholds in GSM networks. *International Journal of Mobile Network Design and Innovation*, 2008. (accepted).

[2] P. Lehtimäki and K. Raivio. Combining measurement data and Erlang-B formula for blocking prediction in GSM networks. In *Proceedings of The 10th Scandinavian Conference on Artificial Intelligence (SCAI)*, Stockholm, Sweden, May 26 - 28 2008.

15.4 Learning from environmental data

Mika Sulkava, Jaakko Hollmén

Data analysis methods play an important role in increasing our knowledge of the environment as the amount of data measured from the environment increases. Gaining an insight into the condition of the environment and the assessment of the its future development under the present and predicted environmental scenarios requires large data sets from long-term monitoring programs. In this project the development of forests in Finland has been studied using data from various forest monitoring programs. In addition, the global changes and drivers of the CO_2 exchange of forests have been studied based on eddy covariance data from a high number of sites around the world.

The work in this project includes collaboration with a high number of parties. During 2006–2007, there has been cooperation with two research units of the Finnish Forest Research Institute, University of Antwerp, and numerous researchers in the carbon cycling community all around the world. The latest journal contributions are joint work of a team of more than a dozen researchers from nine countries in three continents.

Plant nutrients play an integral role in the physiological and biochemical processes of forest ecosystems. The effects of nitrogen and sulfur depositions on coniferous forests have been studied using the Self-Organizing Map. It was concluded that evidence for deposition-induced changes in needles has clearly decreased during the nineties. The results of the effects of the depositions have been presented in conferences [1, 2].

Various environmental factors and past development affect the growth and nutritional composition of tree needles as they are aging. Different regression models have been compared to find out how these effects could be modeled effectively and accurately during the second year of the needles [3]. We found that sparse regression models are well suited for this kind of analysis. They are better for the task than ordinary least squares single and multiple regression models, because they are both easy to interpret and accurate at the same time.

Good quality of analytical measurements techniques is important to ensure the reliability of analyses in environmental sciences. We have combined foliar nutrition data from Finland and results of multiple measurement quality tests from different sources in order to study the effect of measurement quality on conclusions based on foliar nutrient analysis [4, 5]; see Figure 15.3. In particular, we studied the use of weighted linear regression

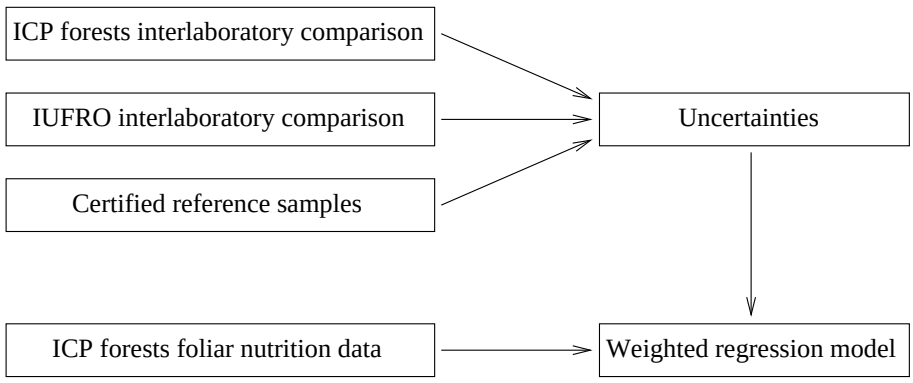


Figure 15.3: Fusion of measurement quality metadata from three different sources and forest nutrition data made it possible to use weighted regression models for trend detection.

models in detecting trends in foliar time series data and showed that good precision of the measurement techniques may decrease the time needed to detect statistically significant trends in environmental time series by several years.

The dependencies between the atmospheric CO_2 exchange of the world's forests and different environmental factors and between the annual radial growth of coniferous trees and environment and properties of the trees have been studied since 2006. First results concerning the significance of photosynthesis in differences between yearly CO_2 exchange have been published lately [6, 7]. Also, the effects of nitrogen deposition on CO_2 exchange in forests have been studied [8].

Finally, the effects of environmental conditions on radial growth of trees has been studied. Methods for automatic detection of the onset and cessation of radial growth [9] and for model selection and estimation based on expert knowledge [10] have been developed.

References

[1] S. Luyssaert, M. Sulkava, H. Raitio, J. Hollmén, and P. Merilä. Is N and S deposition altering the mineral nutrient composition of Norway spruce and Scots pine needles in Finland? In Johannes Eichhorn, editor, *Proceedings of Symposium: Forests in a Changing Environment – Results of 20 years ICP Forests Monitoring*, pages 80–81, Göttingen, Germany, October 2006.

[2] Päivi Merilä, John Derome, Sebastiaan Luyssaert, Mika Sulkava, Jaakko Hollmén, Kaisa Mustajärvi, and Pekka Nöjd. How are N and S in deposition, in percolation water and in upper soil layers reflected in chemical composition of needles in Finland? In *Book of abstracts of Seminar on forest condition monitoring and related studies in northern Europe under the Forest Focus and ICP Forests programmes*, Vantaa, Finland, November 2007.

[3] Mika Sulkava, Jarkko Tikka, and Jaakko Hollmén. Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. *Ecological Modelling*, 191(1):118–130, January 2006.

[4] Mika Sulkava. Modeling how varying data quality affects the ability to detect trends in environmental time series. In Veli Mäkinen, Greger Lindén, and Hannu Toivonen, editors, *Summer School on Algorithmic Data Analysis (SADA 2007) and Annual Hecse Poster Session, Abstract proceedings*, volume B-2007-4 of *Series of Publications B*, page 104, Helsinki, Finland, May/June 2007. University of Helsinki, Department of Computer Science, Helsinki University Printing House.

[5] Mika Sulkava, Sebastiaan Luyssaert, Pasi Rautio, Ivan A. Janssens, and Jaakko Hollmén. Modeling the effects of varying data quality on trend detection in environmental monitoring. *Ecological Informatics*, 2(2):167–176, June 2007.

[6] S. Luyssaert, I. A. Janssens, M. Sulkava, D. Papale, A. J. Dolman, M. Reichstein, T. Suni, J. Hollmén, T. Vesala, D. Lousteau, B. Law, and E. J. Moors. Photosynthesis drives interannual variability in net carbon-exchange of pine forests at different latitudes. In *Proceedings of the Open Science Conference on the GHG Cycle in the Northern Hemisphere*, pages 86–87, Sissi-Lassithi, Greece, November 2006. CarboEurope, NitroEurope, CarboOcean, and Global Carbon Project, Max-Planck-Institute for Biogeochemistry.

[7] S. Luyssaert, I. A. Janssens, M. Sulkava, D. Papale, A. J. Dolman, M. Reichstein, J. Hollmén, J. G. Martin, T. Suni, T. Vesala, D. Lousteau, B. E. Law, and E. J. Moors. Photosynthesis drives anomalies in net carbon-exchange of pine forests at different latitudes. *Global Change Biology*, 13(10):2110–2127, October 2007.

[8] S. Luyssaert, I. Inglima, R. Ceulemans, P. Ciais, A. J. Dolman, J. Grace, J. Hollmén, B. E. Law, G. Matteucci, D. Papale, S. L. Piao, M. Reichstein, E.-D. Schulze, M. Sulkava, J. Tang, and I. A. Janssens. Unravelling nitrogen deposition effects on carbon cycling in forests. *Eos, Transactions, American Geophysical Union*, 88(52), December 2007. Fall Meeting Supplement, Abstract B32B-02.

[9] Mika Sulkava, Harri Mäkinen, Pekka Nöjd, and Jaakko Hollmén. CUSUM charts for detecting onset and cessation of xylem formation based on automated dendrometer data. In Ivana Horová and Jiří Hřebíček, editors, *TIES 2007 – 18th annual meeting of the International Environmetrics Society, Book of Abstracts*, page 111, Mikulov, Czech Republic, August 2007. The International Environmetrics Society, Masaryk University.

[10] Jaakko Hollmén. Model selection and estimation via subjective user preferences. In Vincent Corruble, Masayuki Takeda, and Einoshin Suzuki, editors, *Discovery Science: 10th International Conference, DS 2007, Proceedings*, volume 4755 of *Lecture Notes in Artificial Intelligence*, pages 259–263, Sendai, Japan, October 2007. Springer-Verlag.

15.5 Parsimonious signal representations in data analysis

Jarkko Tikka, Jaakko Hollmén, Timo Similä

The objective in data analysis is to find unsuspected and practical information from large observational data sets and to represent it in a comprehensible way. While utility is a natural starting point for any analysis, understandability often remains a secondary goal. A lot of input variables are available for a model construction in many cases. For instance, in the analysis of microarray data the number of input variables may be tens of thousands. It is impossible to evaluate all the possible combinations of input variables in a reasonable time. In this research, improved understandability of data-analytic models is sought by investigating sparse signal representations that are learned automatically from data. Naturally, the domain expertise is useful in many cases in validation of results, but it may also be biased by established habits and, thus, prevent making novel discoveries.

In a time series context, parsimonious modeling techniques can be used in estimating a sparse set of autoregressive variables for time series prediction [7]. We presented a filter approach to the prediction: first we selected a sparse set of inputs using computationally efficient linear models and then the selected inputs were used in the nonlinear prediction model. Furthermore, we quantified the importance of the individual input variables in the prediction. Based on experiments, our two-phase modeling strategy yielded accurate and parsimonious prediction models giving insight to the original problem.

The problem of estimating sparse regression models in a case of multi-dimensional input and output variables has been investigated in [4]. We proposed a forward-selection algorithm called multiresponse sparse regression (MRSR) that extends the Least Angle Regression algorithm (LARS) [1]. The algorithm was also applied to the task of selecting relevant pixels from images in multidimensional scaling of handwritten digits. The MRSR algorithm was presented in a more general framework in [5]. In addition, experimental comparisons showed the strengths of MRSR against some other input selection methods. The input selection problem for multiple response linear regression was formulated as a

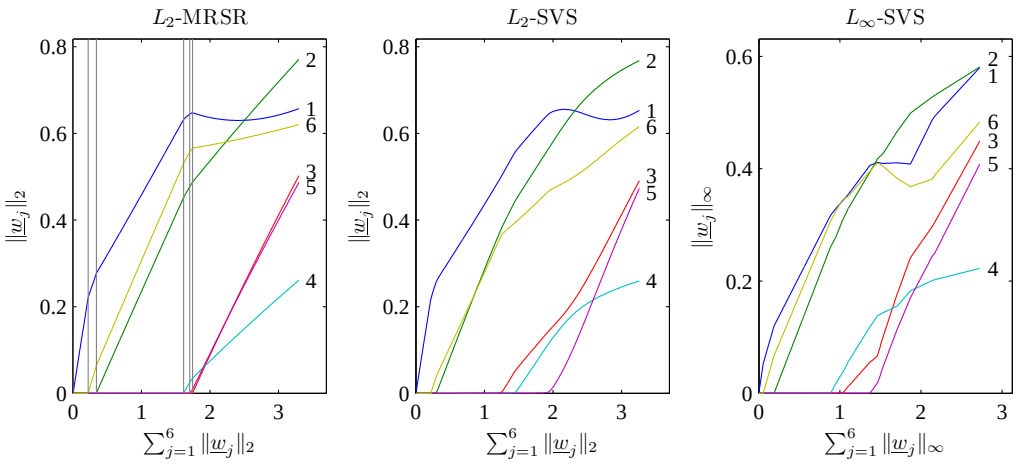


Figure 15.4: Solution paths of the importance factors of input variables. In the subfigure on the left panel, vertical lines indicate the breakpoints of the MRSR algorithm, i.e the points where a new input variable is added to the subset of selected input variables. All the solution paths end to the ordinary least square solution.

convex optimization problem to minimize the error sum of squares subject to a sparsity constraint in [6]. The proposed simultaneous variable selection (L_2-SVS) method is related to L_∞-SVS method [10]. We also reported an efficient algorithm to follow the solution path as a function of the constraint parameter. In Figure 15.4, the solution paths of MRSR, L_2-SVS, and L_∞-SVS are illustrated using a data set, which includes six input variables. The most important inputs are x_2, x_1, and x_6 according to all the three methods. The multiresponse sparse regression is studied further in [2, 3].

The artificial neural networks are an appropriate choice to model dependencies in nonlinear regression problems, since they are capable to approximate a wide class of functions very well. A disadvantage of neural networks is their black-box characteristics. We have developed input selection algorithms for radial basis function (RBF) networks in order to improve their interpretability [8, 9]. A backward selection algorithm (SISAL-RBF), which removes input variables sequentially from the network based on the significance of the individual regressors, was suggested in [9]. The calculation of ranking of inputs is based on partial derivatives of the network. Only 15% of the available inputs were selected by the SISAL-RBF without sacrificing prediction accuracy at all in the case of real world data set [9]. In [8], each input dimension was weighted and a sparsity constraint was imposed on the sum of the weights. The resulting constrained cost function was optimized with respect to the weights and other parameters using alternating optimization approach. The optimum weights describe the relative importance of the input variables. Applications to both simulated and benchmark data produced competitive results.

References

[1] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2): 407–499, 2004.

[2] T. Similä. Majorize-minimize algorithm for multiresponse sparse regression. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Vol. II, pp. 553–556, Honolulu, HI, USA, April 2007.

[3] T. Similä. *Advances in variable selection and visualization methods for analysis of multivariate data*. PhD Thesis, Helsinki University of Technology, 2007.

[4] T. Similä and J. Tikka. Multiresponse sparse regression with application to multidimensional scaling. *Proceedings of the 15th International Conference on Artificial Neural Networks (ICANN 2005)*, Vol. 3967 (part II) of Lecture Notes in Computer Science, Springer, pp. 97–102, Warsaw, Poland, September, 2005.

[5] T. Similä and J. Tikka. Common subset selection of inputs in multiresponse regression. *Proceedings of the 19th International Joint Conference on Neural Networks (IJCNN 2006)*, pp. 1908–1915, Vancouver, Canada, July, 2006.

[6] T. Similä and J. Tikka. Input selection and shrinkage in multiresponse linear regression. *Computational Statistics & Data Analysis*, 52(1): 406–422, 2007.

[7] J. Tikka and J. Hollmén. Sequential input selection algorithm for long-term prediction of time series. *Neurocomputing*. Accepted for publication.

[8] J. Tikka. Input selection for radial basis function networks by constrained optimization. *Proceedings of the 17th International Conference on Artificial Neural Networks (ICANN 2007)*, Vol. 4668 of Lecture Notes in Computer Science, Springer, pp. 239–248, Porto, Portugal, September, 2007.

[9] J. Tikka and J. Hollmén. Selection of important input variables for RBF network using partial derivatives. *Proceedings of the 16th European Symposium on Artificial Neural Networks (ESANN 2008)*. In press.

[10] B.A. Turlach, W.N. Venables, and S.J. Wright. Simultaneous variable selection Technometrics, 47(3): 349–363, 2005.

Chapter 16

Time series prediction

Amaury Lendasse, Francesco Corona, Antti Sorjamaa, Elia Liitiäinen, Tuomas Kärnä, Yu Qi, Emil Eirola, Yoan Miché, Yongnang Ji, Olli Simula

16.1 Introduction

Amaury Lendasse

What is Time series prediction? Time series prediction (TSP) is a challenge in many fields. In finance, experts forecast stock exchange courses or stock market indices; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the load of the following day. The common point to their problems is the following: how can one analyse and use the past to predict the future? Many techniques exist: linear methods such as ARX, ARMA, etc., and nonlinear ones such as artificial neural networks. In general, these methods try to build a model of the process. The model is then used on the last values of the series to predict the future values. The common difficulty to all the methods is the determination of sufficient and necessary information for an accurate prediction.

A new challenge in the field of time series prediction is the Long-Term Prediction: several steps ahead have to be predicted. Long-Term Prediction has to face growing uncertainties arising from various sources, for instance, accumulation of errors and the lack of information.

Our contributions in TSP research. The TSP group is a new research group. It has been created in 2004. A notable achievement has been the organization of the first European Symposium on Time Series Prediction (ESTSP'07) on February 2007 in Helsinki. (`http://www.estsp.org`, [1]). For this symposium, a time series competition has been organized and a benchmark has been created.

In the reporting period 2006 - 2007, TSP research has been established as a new project in the laboratory. Nevertheless, TSP research has already been extended to a new direction: "Chemoinformatics".

This Chapter starts by introducing some theoretical advances undertaken during the reporting period, including the presentation of the ESTSP´07 competition. Also the problem of input selection for TSP is reported. The applications range includes Chemoinformatics.

16.2 European Symposium on Time Series Prediction

Amaury Lendasse and Antti Sorjamaa

Time series forecasting is a challenge in many fields. In finance, experts forecast stock exchange courses or stock market indices; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the load of the following day. ESTSP 2007 was a unique opportunity for researcher from Statistics, Neural Networks, Machine Learning, Control and Econometrics to share their knowledge in the field of Time Series Prediction.

The common point to their problems is the following: how can one analyse and use the past to predict the future?

Many techniques exist for the approximation of the underlying process of a time series: linear methods such as ARX, ARMA, etc. , and nonlinear ones such as artificial neural networks. In general, these methods try to build a model of the process. The model is then used on the last values of the series to predict the future values. The common difficulty to all the methods is the determination of sufficient and necessary information for an accurate prediction.

A new challenge in the field of time series prediction is the Long-Term Prediction: several steps ahead have to be predicted. Long-Term Prediction has to face growing uncertainties arising from various sources, for instance, accumulation of errors and the lack of information to predict the future values.

Papers were presented orally (single track).

The following is a non-exhaustive list of machine learning, computational intelligence and artificial neural networks topics covered during the ESTSP conferences:

- Short-term prediction
- Long-term prediction
- Econometrics
- Nonlinear models for Time Series Prediction
- Time Series Analysis
- Prediction of non-stationary Time Series
- System Identification
- System Identification for control
- Feature (variable or input) Selection for Time Series
- Selection of Exogenous (external) variables

The goal of the competition is the prediction of the 50 next values (or more) of the time series. The evaluation of the performance was done using the MSE obtained from the prediction of both the 15 and the 50 next values.

So far, there are now 74 values available and the results can be found in http://www.cis.hut.fi/projects/tsp/ESTSP/. In the following figure the predictions of all the competition participants are plotted in blue. In red and in the table below are shown the real values so far.

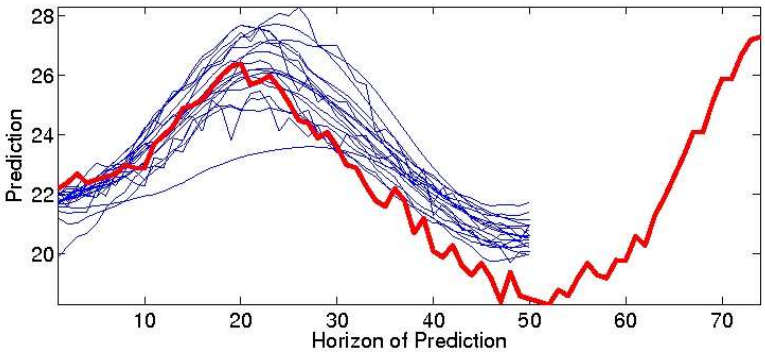


Figure 16.1: The ESTSP Benchmark.

16.3 Methodology for long-term prediction of time series

Amaury Lendasse, Yu Qi, Yoan Miché and Antti Sorjamaa

The time series prediction problem is the prediction of future values based on the previous values and the current value of the time series (see Equation 16.1).

$$\hat{y}_{t+1} = f_1(y_t, y_{t-1}, ..., y_{t-M+1}). \tag{16.1}$$

The previous values and the current value of the time series are used as inputs for the prediction model. One-step ahead prediction is needed in general and is referred as Short-Term Prediction. But when multi-step ahead predictions are needed, it is called Long-Term Prediction problem.

Unlike the Short-Term time series prediction, the Long-Term Prediction is typically faced with growing uncertainties arising from various sources. For instance, the accumulation of errors and the lack of information make the prediction more difficult. In Long-Term Prediction, performing multiple steps ahead prediction, there are several alternatives to build models. Two variants of prediction strategies are studied and compared [2]: the Direct (see Equation 16.2) and the Recursive Prediction Strategies (see Equation 16.1).

$$\hat{y}_{t+k} = f_k(y_t, y_{t-1}, ..., y_{t-M+1}). \tag{16.2}$$

16.4 Nonparametric noise estimation

Elia Liitiäinen, Francesco Corona, Emil Eirola and Amaury Lendasse

The residual variance estimation problem (or Nonparametric noise Estimation) is well-known in machine learning and statistics under various contexts. Residual variance estimation can be viewed as the problem of estimating the variance of the part of the output that cannot be modelled with the given set of input variables. This type of information is valuable and gives elegant methods to do model selection. While there exist numerous applications of residual variance estimators to supervised learning, time series analysis and machine learning, it seems that a rigorous and general framework for analysis is still missing. For example, in some publications the theoretical model assumes additive noise and independent identically distributed (iid) variables. The principal objective of our work is to define such a general framework for residual variance estimation by extending its formulation to the non-iid case. The model is chosen to be realistic from the point of view of supervised learning. Secondly, we view two well-known residual variance estimators, the Delta test and the Gamma test in the general setting and we discuss their convergence properties. Based on the theoretical achievements, our general approach seems to open new directions for future research and it appears of fundamental nature [3]. We have also applied NNE for time series prediction [4].

16.5 Chemoinformatics

Francesco Corona, Elia Liitiäinen, Tuomas Kärnä and Amaury Lendasse

Many analytical problems related to spectrometry require predicting a quantitative variable through a set of measured spectral data. For example, one can try to predict a chemical component concentration in a product through its measured infrared spectrum. In recent years, the importance of such problems in various fields including the pharmaceutical, food and textile industries have grown dramatically. The chemical analysis by spectrophotometry rests on the fast acquisition of a great number of spectral data (several hundred, even several thousands).

In spectrometric problems, one is often faced with databases having more variables (spectra components) than samples; and almost all models use at least as many parameters as the number of input variables. These two problems, colinearity and risk of overfitting, already exist in linear models. However, their effect may be even more dramatic when nonlinear models are used (there are usually more parameters than in linear models, and the risk of overfitting is higher). In such high-dimensional problems, it is thus necessary to use a smaller set of variables than the initial one. We have proposed methods to select spectral variables by using concepts from information theory:

- the measure of mutual information [5].
- the measure of topological relevance on the Self-Organizing Map [6]
- the Functional Data Analysis (FDA) [7]
- Nonparametric Noise Estimation [8]

One particular application has been studied in the field of Oil Production.

In this industrial application, there has been applied process data from Neste Oil Oyj. The aim has been to get new empirical modelling tools, which are based on information technology. The outcome has been emphasized on tools, which are suitable in fast data mining from large data sets. The test cases have included:

- Analysis of instrumental data, on-line monitoring data and quality data
- Non-linear processes
- Identification of delays between stages in industrial processes
- Robust variable selection methods

Analysis of instrumental data, on-line monitoring data and quality data The case has been progressed using a real process data set having 13000 on-line samples (time points) and over a thousand variables. The variables contained different blocks: Z (NIR), X (Process variables) and Y (Quality of end product).

References

[1] Amaury Lendasse. European Symposium on Time Series Prediction, ESTSP'07, Amaury Lendasse editor, ISBN 978-951-22-8601-0.

[2] Amaury Sorjamaa, Jin Hao, Nima Reyhani, Yongnang Ji and Amaury Lendasse, Methodology for Long-term Prediction of Time Series Neurocomputing (70), October 16-18, 2007, pp. 2861-2869

[3] E. Liitiäinen, Francesco Corona and Amaury Lendasse, Non-parametric Residual Variance Estimation in Supervised Learning IWANN 2007, International Work-Conference on Artificial Neural Networks, San Sebastian (Spain), June 20-22, Springer-Verlag, Lecture Notes in Computer Science, 2007, pp. 63-71.

[4] Elia Liitiäinen and A. Lendasse, Variable Scaling for Time Series Prediction: Application to the ESTSP'07 and the NN3 Forecasting Competitions IJCNN 2007, International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007.

[5] Fabrice Rossi, Amaury Lendasse, Damien François, Vincent Wertz and Michel Verleysen. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling, Chemometrics and Intelligent Laboratory Systems, Volume 80, Issue 2, 15 February 2006, pages 215-226.

[6] Francesco Corona, Satu-Pia Reinikainen, Kari Aaljoki, Anniki Perkkio, Elia Liitiäinen, Roberto Baratti and Amaury Lendasse and Olli Simula. Wavelength selection using the measure of topological relevance on the Self-Organizing Map, Journal of Chemometrics, submitted and accepted in 2007.

[7] Tuomas Kärnä and Amaury Lendasse, Gaussian fitting based FDA for chemometrics, IWANN'07, International Work-Conference on Artificial Neural Networks, San Sebastian, Spain, June 20-22 , 86–193, 2007.

[8] Amaury Lendasse and Francesco Corona Optimal Linear Projection based on Noise Variance Estimation - Application to Spectrometric Modeling SSC10, 10th Scandinavian Symposium on Chemometrics, Lappeenranta (Finland) June 11-15, 2007

Individual projects

A. Approximation of an input data item by a linear mixture of SOM models

Teuvo Kohonen

The purpose of this work was to extend the use of the SOM by showing that instead of a single 'winner' model, one can approximate the input data item more accurately by means of a set of *several models* that *together* define the input data item more accurately. It shall be emphasized that we do not mean 'k winners' that are rank-ordered according to their matching. Instead, the input data item is approximated by an *optimized linear mixture of the models, using a nonlinear constraint*, which will be shown to provide an improved description of it.

Consider the n-dimensional SOM models $\mathbf{m}_i, i = 1, 2, \ldots, p$, where p is the number of nodes in the SOM. Their general linear mixture is written as

$$k_1\mathbf{m}_1 + k_2\mathbf{m}_2 + \ldots + k_p\mathbf{m}_p = \mathbf{M}\,\mathbf{k} \ , \tag{I.1}$$

where the k_i are scalar-valued weighting coefficients, $\mathbf{k}$ is the p-dimensional column vector formed of them, and $\mathbf{M}$ is the matrix with the $\mathbf{m}_i$ as its columns. Now $\mathbf{M}\,\mathbf{k}$ shall be the *estimate* of some input vector $\mathbf{x}$. The vectorial fitting error is then

$$\mathbf{e} = \mathbf{M}\,\mathbf{k} - \mathbf{x} \ . \tag{I.2}$$

Our aim is to minimize the norm of $\mathbf{e}$ in the sense of least squares. However, a special constraint must then be taken into account.

Fitting with the nonnegativity constraint

Much attention has recently been paid to least-squares problems where the fitting coefficients are constrained to *nonnegative values*. Such a constraint is natural, when the *negatives* of the items have no meaning, for instance, when the input item consists of statistical indicators that can have only nonnegative values, or is a weighted word histogram of a document. In these cases at least, the constraint contains additional information that is expected to make the fits more meaningful.

The lsqnonneg function

The present fitting problem belongs to the broader category of *quadratic programming* or *quadratic optimization*, for which numerous methods have been developed in recent years. A much-applied one-pass algorithm is based on the *Kuhn-Tucker theorem* (Lawson & Hanson, 1974), but it is too involved to be reviewed here in full. Let it suffice to mention that it has been implemented in Matlab as the function named the *lsqnonneg*. Below, the variables $\mathbf{k}$, $\mathbf{M}$, and $\mathbf{x}$ must be understood as being defined in the Matlab format. Then we obtain the weight vector $\mathbf{k}$ as

$$\mathbf{k} = \mathrm{lsqnonneg}(\mathbf{M}, \mathbf{x}) \ . \tag{I.3}$$

The *lsqnonneg* function can be computed, and the result will be meaningful, for an *arbitrary rank* of the matrix $\mathbf{M}$. Nonetheless it has to be admitted that there exists a rare theoretical case where the optimal solution is not *unique*. This case occurs, if some of the $\mathbf{m}_i$ *in the final optimal mixture* are *linearly dependent*. In practice, if the input data items to the SOM are stochastic, the probability for the optimal solution being not unique is negligible. At any rate, the locations of the nonzero weights are unique even in this case!

Description of a document by a linear mixture of SOM models

The following analysis applies to most of the SOM applications. Here it is exemplified by textual data bases.

In text analysis, one possible task is to find out whether a text comes from different sources, whereupon its word histogram is expected to be a linear mixture of other known histograms.

The text corpus used in this experiment was taken from a collection published by the Reuters corporation. No original documents were made available; however, Lewis et al. (2004), who have prepared this corpus for benchmarking purposes, have preprocessed the textual data, removing the stop words and reducing the words into their stems. Our work commenced with the ready word histograms. J. Salojärvi from our laboratory selected a 4000-document subset from this preprocessed corpus, restricting only to such articles that were assigned to one of the following classes:

1. Corporate-Industrial.
2. Economics and Economic Indicators.
3. Government and Social.
4. Securities and Commodities Trading and Markets.

There were 1000 documents in each class. Salojärvi then picked up those 1960 words that appeared at least 200 times in the selected texts. In order to carry out *statistically independent experiments*, a few documents were set aside for testing. The 1960-dimensional word histograms were weighted by factors used by Manning and Schütze [3]. Using the weighted word histograms of the rest of the 4000 documents as input, a 2000-node SOM was constructed.

Fig. I.2 shows the four distributions of the hits on the SOM, when the input items from each of the four classes were applied separately to the SOM. It is clearly discernible that the map is *ordered*, i.e., the four classes of documents are segregated to a reasonable accuracy, and the mappings of classes 1, 3, and 4 are even singly connected, in spite of their closely related topics.

Fig. I.3 shows a typical example, where a linear mixture of SOM models was fitted to a new, unknown document. The values of the weighting coefficients k_i in the mixture are shown by dots with relative intensities of color in the due positions of the SOM models. It is to be emphasized that this fitting procedure also defines the optimal *number* of the nonzero coefficients. In the experiments with large document collections, this number was usually very small, less than a per cent of the number of models.

When the models fall in classes that are known a priori, the weight of a model in the linear mixture also indicates the *weight of the class label associated with that model*. Accordingly, by summing up the weights of the various types of class labels one then obtains the *class-affiliation* of the input with the various classes.

References

[1] C.L. Lawson and R.J. Hanson, *Solving Least-Squares Problems*, Englewood Cliffs, NJ: Prentice-Hall, 1974

[2] D.D. Lewis, Y. Yang, T.G. Rose, and T. Li, "RCV1: A new benchmark collection for text categorization research," *J. Mach. Learn. Res.* vol. 5, pp.361-397, 2004.

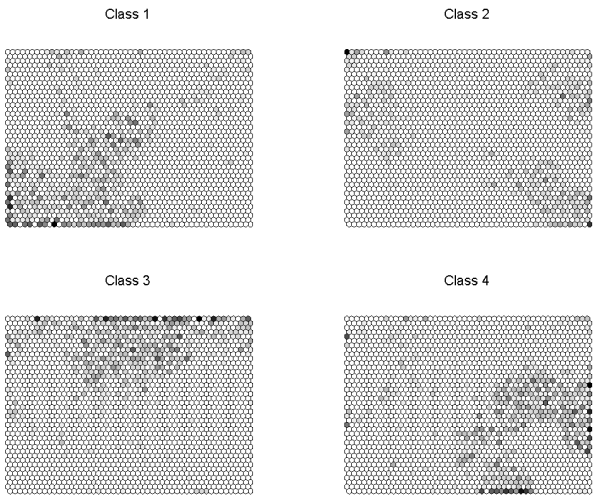


Figure I.2: Mapping of the four Reuters document classes onto the SOM. The densities of the "hits" are shown by shades of gray.

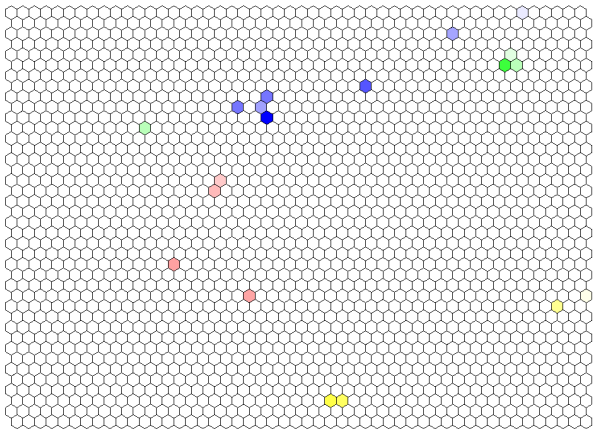


Figure I.3: A linear mixture of SOM models fitted to a new, unknown document. The weighting coefficients k_i in the mixture are shown by using a coloring with a relative saturation of the due models. Red: Class 1. Green: Class 2. Blue: Class 3. Yellow: Class 4.

[3] C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press, 1999.

B. Independent variable group analysis

Krista Lagus, Antti Honkela, Jeremias Seppä, Paul Wagner

Independent variable group analysis (IVGA) [1, 2] is a principle for grouping observed input variables so that mutual dependences between variables are strong within a group and weak between groups.

In problems with a large number of diverse observations there are often groups of input variables that have strong mutual dependences within the group but which can be considered practically independent of the input variables in other groups. It can be expected that the larger the problem domain, the more independent groups there are. Estimating a model for each independent group separately produces a more compact representation than applying the model to the whole set of variables. Compact representations are computationally beneficial and, moreover, offer better generalization.

Usually such variable grouping is performed by a domain expert, prior to modeling with automatic, adaptive methods. As expert knowledge may be unavailable, or expensive and time-consuming, automating the task can considerably save resources. The IVGA is a practical, efficient and general approach for obtaining compact representations that can be regarded as sparse codes, as well. Software packages implementing all the presented algorithms are available at `http://www.cis.hut.fi/projects/ivga/`.

The IVGA project is a collaboration with Dr. Esa Alhoniemi (University of Turku) and Dr. Harri Valpola (Helsinki University of Technology, Laboratory of Computational Engineering).

The IVGA algorithm

Any IVGA algorithm consists of two parts, (1) grouping of variables, and (2) construction of an independent model for each variable group. A variable grouping is obtained by comparing models under different groupings using a suitable cost function. In principle any model can be used, if the necessary cost function is derived for the model family.

A practical grouping algorithm for implementing the IVGA principle was first presented in [1]. The method used vector quantizers (VQs) learned with variational Bayesian methods [3] to model the individual groups.

In more recent work [2] we have shown that the variational Bayesian approach is approximately equivalent to minimizing the mutual information or multi-information between the groups. Additionally the modelling algorithm was extended to a finite mixture model that can handle mixed data consisting of both real valued and nominal variables.

Agglomerative grouping algorithm

In addition to the regular combinatorial grouping algorithm corresponding to regular clustering, we have developed an agglomerative IVGA (AIVGA) algorithm for hierarchical grouping of variables [4, 5].

The agglomerative algorithm provides a hierarchical grouping of the variables by starting from singleton groups and merging them iteratively. This both provides a hierarchical view of the variable dependencies as well as a simple greedy deterministic algorithm for solving the ordinary IVGA problem. Experiments reported in [5] show that this method can greatly simplify application of IVGA to practical problems.

Application to computational biology

In [6], IVGA was used to find independent groups of genes or gene regulatory modules from measurements of transcription factor protein binding to different genes in the DNA. The independence of these modules was verified by studying the estimated mutual information of gene expression measurements from mutant organisms with different genes in the discovered modules knocked out. The modules found by IVGA were found to be more meaningful than those discovered by conventional clustering methods.

References

[1] K. Lagus, E. Alhoniemi, and H. Valpola, "Independent variable group analysis," in *Proc. Int. Conf. on Artificial Neural Networks - ICANN 2001*, ser. LNCS, vol. 2130. Vienna, Austria: Springer, 2001, pp. 203–210.

[2] E. Alhoniemi, A. Honkela, K. Lagus, J. Seppä, P. Wagner, and H. Valpola, "Compact modeling of data using independent variable group analysis," *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 1762–1776, 2007.

[3] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. Jordan, Ed. Cambridge, MA, USA: The MIT Press, 1999, pp. 105–161.

[4] A. Honkela, J. Seppä, and E. Alhoniemi, "Agglomerative independent variable group analysis," in *Proc. 15th European Symposium on Artificial Neural Networks (ESANN 2007)*, Bruges, Belgium, 2007, pp. 55–60.

[5] A. Honkela, J. Seppä, and E. Alhoniemi, "Agglomerative independent variable group analysis," *Neurocomputing*, 2008, doi:10.1016/j.neucom.2007.11.024.

[6] J. Nikkilä, A. Honkela, and S. Kaski, "Exploring the independence of gene regulatory modules," in *Proc. Workshop on Probabilistic Modeling and Machine Learning in Structural and Systems Biology*, J. Rousu, S. Kaski, and E. Ukkonen, Eds., Tuusula, Finland, 2006, pp. 131–136.

C. Analysis of discrete diffusion scale-spaces

Ramūnas Girdziušas

Taking averages of observations is the most basic method to make inferences in the presence of uncertainty. In late 1980's, this simple idea has been extended to the principle of *successively average less where the change is faster*, and applied to the problem of revealing a signal with jump discontinuities in additive noise.

Successive averaging results in a family of signals with progressively decreasing amount of details, which is called the *scale-space* and further conveniently formalized by viewing it as a solution to a certain diffusion-inspired evolutionary partial differential equation (PDE). Such a model is known as the *diffusion scale-space*.

Example of linear and nonlinear diffusion scale-spaces are shown in Fig. I.4. Diffusion scale-spaces possess two long-standing problems: (i) *model analysis* which aims at establishing stability and guarantees that averaging does not distort important information, and (ii) *model selection*, such as identification of the optimal scale (diffusion stopping time) given an initial noisy signal and an incomplete model.

This thesis studies both problems in the discrete space and time. Such a setting has been strongly advocated by Lindeberg (1991) and Weickert (1996) among others. The focus of the model analysis part is on necessary and sufficient conditions which guarantee that a discrete diffusion possesses the scale-space property in the sense of sign variation diminishing. Connections with the total variation diminishing and the open problem in a multivariate case are discussed too.

Considering the model selection, the thesis unifies two optimal diffusion stopping principles: (i) the time when the Shannon entropy-based Liapunov function of Sporring and Weickert (1999) reaches its steady state, and (ii) the time when the diffusion outcome has the least correlation with the noise estimate, contributed by Mrazek and Navara (2003). Both ideas are shown to be particular cases of the marginal likelihood inference, which is also communicated in [1]. Moreover, the suggested formalism provides first principles behind such criteria, and removes a variety of inconsistencies. It is suggested that the outcome of the diffusion should be interpreted as a certain expectation conditioned on the initial signal of observations instead of being treated as a random sample or probabilities.

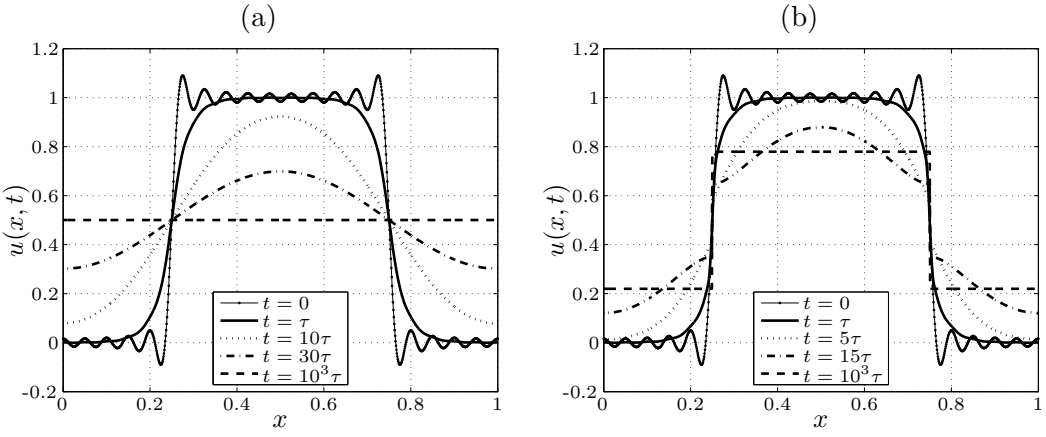


Figure I.4: Example of linear (a) and nonlinear (b) diffusion scales-spaces. They consist of signals which are indexed by the scale value t.

This removes the need to normalize signals, and it also better justifies application of the correlation criterion.

As an example, the following improvement to the existing results can be mentioned. Let the generalized Laplacian $\mathbf{B} \in \mathbb{R}^{n \times n}$ be defined as:

$$\mathbf{B} \equiv - \begin{pmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix} \begin{pmatrix} b_1 & & \\ & \ddots & \\ & & b_{n+1} \end{pmatrix} \begin{pmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}^T , \tag{I.4}$$

where the first matrix is bidiagonal and it contains n rows and $n+1$ columns, whereas the second matrix is diagonal of size $n+1$. Undesignated elements are zeroes.

Theorem 1 *Let $Q_{k,n}$ denote the totality of $\binom{n}{k}$ increasing sequences of integers, each taken from $\{1, 2, \ldots, n\}$ and being of length k. Given any sequence $\boldsymbol{\omega} \in Q_{k,n}$, divide it into r groups of connected indices:*

$$\boldsymbol{\omega} = \{\underbrace{\omega_1, \ldots, \omega_{\nu_1}}_{1st\ group}, \underbrace{\omega_{\nu_1+1}, \ldots, \omega_{\nu_2}}_{2nd\ group}, \ \ldots\ , \underbrace{\omega_{\nu_{r-1}+1}, \ldots, \omega_k}_{r\text{-}th\ group}\} . \tag{I.5}$$

A particular example of $Q_{3,20}$ such as $7, 11, 12$ would produce two groups.

1. *The matrix $(\mathbf{I} - \tau\mathbf{B})^{-1}$ is positive definite if and only if*

$$1 + \sum_{k=1}^{p} \tau^k \sum_{\boldsymbol{\omega} \in Q_{k,p}} \prod_{s=1}^{r} \sum_{i=\nu_{s-1}}^{\nu_s+1} \prod_{\substack{j=\nu_s-1 \\ j \neq i}}^{\nu_s+1} b_j > 0, \quad \textit{for all } p = 1, \ldots, n. \tag{I.6}$$

   In particular, the constraint is satisfied if $b_i > 0$ for all $i = 2, \ldots, n$.

2. *The matrix $\mathbf{I} + \tau\mathbf{B}$ is positive definite if and only if*

$$1 + \sum_{k=1}^{p} \tau^k \sum_{\boldsymbol{\omega} \in Q_{k,p}} \prod_{s=1}^{r} \sum_{i=\nu_{s-1}}^{\nu_s+1} \prod_{\substack{j=\nu_s-1 \\ j \neq i}}^{\nu_s+1} (-b_j) > 0, \quad \textit{for all } p = 1, \ldots, n. \tag{I.7}$$

   In particular, the constraint is satisfied if $0 \leq b_2 \leq \tau^{-1}$, $0 \leq b_{i-1} + b_i \leq \tau^{-1}$ for all $i = 3, \ldots, n$, and $0 \leq b_n \leq \tau^{-1}$.

The proof with a geometric description can be found in [3]. All known previous characterizations either include determinantal quantities, or provide only sufficient conditions such as the case of positive diagonal elements with a diagonal dominance, which follows from Gershgorin's circles. Theorem 1 has been related to the scale-space analysis in [2].

References

[1] R. Girdziušas and J. Laaksonen. How marginal likelihood inference unifies entropy, correlation and snr-based stopping in nonlinear diffusion scale-spaces. In I. S. Kweon Y. Yagi, S. B. Kang and H. Zha, editors, *Proc. of 8th Asian Conf. on Computer Vision*, volume 4843 of *Lecture Notes in Computer Science*, pages 811–820, 2007.

[2] R. Girdziušas. *Stability and Inference in Discrete Diffusion Scale-Spaces.* Doctoral thesis, Helsinki University of Technology, 2008.

[3] R. Girdziušas and J. Laaksonen. When is a discrete diffusion a scale-space? In A. Shashua D. Metaxas, B. C. Vemuri and H. Shum, editors, *Proc. of 11th IEEE Int. Conf. on Computer Vision*, page 6. IEEE, 2007.

D. Feature selection for steganalysis

Yoan Miche, Amaury Lendasse, Patrick Bas and Olli Simula

Steganography has been known and used for a very long time, as a way to exchange information in an unnoticeable manner between parties, by embedding it in another, apparently innocuous, document. For example, during the 80's, Margaret Thatcher decided to have each word processor of the government's administration members changed with an unique word spacing for each, giving a sort of "invisible signature" to documents. This was done to prevent the continuation of sensitive government information leaks.

Nowadays steganographic techniques are mostly used on digital contents. The online newspaper, Wired News, reported in one of its articles on steganography that several steganographic contents have been found on web-sites with very large image database such as eBay.

Most of the time research about steganography is not as much to hide information, but more to detect that there is hidden information. This "reverse" part of the steganography is called steganalysis and is specifically aimed at making the difference between genuine documents, and steganographied – called stego – ones. Consequently, steganalysis can be seen as a classification problem where the goal is to build a classifier able to distinguish these two sorts of documents.

During the steganographic process, a message is embedded in an image so that it is as undetectable as possible. Basically, it uses several heuristics in order to guarantee that the statistics of the stego content (the modified image) are as close as possible to the statistics of the original one. Afterwards, steganalysis techniques classically use features extracted from the analyzed image and an appropriately trained classifier to decide whether the image is genuine or not.

In our work, a widely used and known set of 193 image features has been used. Theses features consider statistics of JPEG compressed images such as histograms of DCT coefficients for different frequencies, histograms of DCT coefficients for different values, global histograms, blockiness measures and co-occurrence measures. The main purpose of this high number of features is to obtain a model able to detect about any steganographic process.

The usual process in steganalysis is then to train a classifier according to the extracted features. Consequently a set of 193 features for each image of the database is obtained, giving an especially high dimensionality space for classifiers to work on. Earlier research about these high dimensionality spaces has shown that a lot of issues come out when the number of features is as high as this one.

The main idea behind the carried out work [1, 2, 3, 4] is to give insights on proper handling and use of such high dimensionality datasets; indeed, these are very common in the steganography/steganalysis field and users tend not to respect basic principles (for example having a sufficient number of samples regarding the dimensionality of the problem). In the framework of an international thesis co-agreement between the GIPSA-lab in Institut National Polytechnique de Grenoble (France) and ICS laboratory in Helsinki University of technology, Yoan Miche (GIPSA-lab, ICS) along with Patrick Bas (GIPSA-lab) and Amaury Lendasse (ICS), his advisors, as well as Olli Simula (ICS), his supervisor, developed a methodology for handling these datasets; this methodology is used to determine a sufficient number of images for effective training of a classifier in the obtained high-dimensional space, and use feature selection to select most relevant features for the

desired classification. Dimensionality reduction managed to reduce the original 193 features set by a factor of 13, with overall same performance.

By the use of a Monte-Carlo technique on up to 4000 images, it has been shown that such numbers of images are sufficient for stable results when having a set of 193 features extracted from all images. In the experiments, dimensionality reduction managed to reduce the number of required features to 14, while keeping roughly the same classification results. Computational time is thus greatly improved, divided by about 11. Also, further analysis becomes again possible with this low number of features: conclusions and precisions about the steganographic scheme can be inferred from the obtained feature set.

References

[1] Y. Miche and P. Bas and A. Lendasse and O. Simula and C. Jutten, *Avantages de la Sélection de Caractéristiques pour la Stéganalyse*, in GRETSI 2007, Groupe de Recherche et d'Etudes du Traitement du Signal et des Images, Troyes, France, September 11-13 2007.

[2] Y. Miche and P. Bas and A. Lendasse and C. Jutten and O. Simula, *Advantages of Using Feature Selection Techniques on Steganalysis Schemes*, in IWANN'07: International Work-Conference on Artificial Neural Networks, San Sebastian, Spain, June 20-22 2007.

[3] Y. Miche and P. Bas and A. Lendasse and C. Jutten and O. Simula, *Extracting Relevant Features of Steganographic Schemes by Feature Selection Techniques*, in Wacha'07: Third Wavilla Challenge, Saint Malo, France, June 14 2007.

[4] Y. Miche and B. Roue and P. Bas and A. Lendasse, *A Feature Selection Methodology for Steganalysis*, in MRCS06, International Workshop on Multimedia Content Representation, Classification and Security, Istanbul, Turkey, September 11-13 2006.

E. Adaptive committee techniques

Matti Aksela, Jorma Laaksonen, Erkki Oja

Combining the results of several classifiers can improve performance because in the outputs of the individual classifiers the errors are not necessarily overlapping. Also the combination method can be adaptive. The two most important features of the member classifiers that affect the committee's performance are their individual error rates and the diversity of the errors. The more different the mistakes made by the classifiers, the more beneficial the combination of the classifiers can be.

Selecting member classifiers is not necessarily simple. Several methods for classifier diversity have been presented to solve this problem. In [2] a scheme weighting similar errors made in an exponential fashion, the Exponential Error Count method, was found to provide good results. Still, the best selection of member classifiers is highly dependent on the combination method used.

We have experimented with several adaptive committee structures. Two effective methods have been the Dynamically Expanding Context (DEC) and Class-Confidence Critic Combining (CCCC) schemes. The DEC algorithm was originally developed for speech recognition purposes. The main idea is to determine just a sufficient amount of context for each individual segment so that all conflicts in classification results can be resolved. In the DEC committee, the classifiers are initialized and ranked in the order of decreasing performance. Results of the member classifiers are used as a one-sided context for the creation of the DEC rules. Each time a character is input to the system, the existing rules are searched through. If no applicable rule is found, the default decision is applied. If the recognition was incorrect, a new rule is created.

In our CCCC approach the main idea is to try to produce as good as possible an estimate on the classifier's correctness based on its prior behavior for the same character class. This is accomplished by the use of critics that assign a confidence value to each classification. The confidence value is obtained through constructing and updating distribution models of distance values from the classifier for each class in every critic. These distribution models are then used to extract the needed confidence value, based on prior results in addition to the sample being processed. The committee then uses a decision mechanism to produce the final output from the input label information and critic confidence values. In our earlier experiments the adaptive committee structures have been shown to be able to improve significantly on their members' results.

Also classifiers that are adaptive in themselves can be combined using an adaptive committee. Experiments have shown that while making a single classifier adaptive does produce on average the best gains when used alone, the addition of another layer of adaptation, when implemented in a robust fashion, can produce even better results than either method alone [1].

References

[1] Matti Aksela and Jorma Laaksonen. Adaptive combination of adaptive classifiers for on-line handwritten character recognition. *Pattern Recognition Letters*, 28(1):136–143, 2007.

[2] Matti Aksela and Jorma Laaksonen. Using diversity of errors for selecting members of a committee classifier. *Pattern Recognition*, 39(4):608–623, 2006.

Publications of the Adaptive Informatics Research Centre

[1] A. Ajanki, J. Nikkilä, and S. Kaski. Discovering condition-dependent Bayesian networks for gene regulation. In *Proceedings of the Fifth IEEE International Workshop on Genomic Signal Processing and Statistics*, 2007.

[2] M. Aksela and J. Laaksonen. Using diversity of errors for selecting members of a committee classifier. *Pattern Recognition*, 39:608–623, 2006.

[3] M. Aksela and J. Laaksonen. Adaptive combination of adaptive classifiers for handwritten character recognition. *Pattern Recognition Letters*, 28(1):136–143, 2007.

[4] E. Alhoniemi, A. Honkela, K. Lagus, J. Seppä, P. Wagner, and H. Valpola. Compact modeling of data using independent variable group analysis. Technical Report Report E3, Helsinki University of Technology, Espoo, 2006.

[5] E. Alhoniemi, A. Honkela, K. Lagus, J. Seppä, P. Wagner, and H. Valpola. Compact modeling of data using independent variable group analysis. *IEEE Transactions on Neural Networks*, 18(6):1762–1776, 2007.

[6] M. S. C. Almeida, H. Valpola, and J. Särelä. Separation of nonlinear image mixtures by denoising source separation. In *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Signal Separation, ICA 2006*, pages 8–15, 2006.

[7] P. Bas and F. Cayre. Achieving subspace or key security for WOA using natural or circular watermarking. In *ACM Multimedia and Security Workshop 2006*, 2006.

[8] P. Bas and F. Cayre. Natural watermarking: a secure spread spectrum technique for WOA. In *8th Information Hiding Workshop (IH 2006) 2006*, 2006.

[9] P. Bas and J. Hurri. Vulnerability of DM watermarking of non-iid host signals to attacks utilising the statistics of independent components. *IEE Proceedings - Information Security*, 153(3):127–139, 2006.

[10] S. Borisov, A. Ilin, R. Vigário, and E. Oja. Comparison of BSS methods for the detection of alpha-activity components in EEG. In *6th Int. Conf. on Independent Component Analysis and Blind Signal Separation 2006*, pages 430–437, 2006.

[11] G. J. Brown and K. J. Palomäki. Reverberation. In D. Wang and G. J. Brown, editors, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, pages 209–250, USA, 2006. Wiley/IEEE Press.

[12] F. Corona and A. Lendasse. Variable scaling for time series prediction. In *ESTSP 2007, European Symposium on Time Series Prediction, Espoo (Finland)*, pages 69–76, 2007.

[13] F. Corona, E. Liitiäinen, A. Lendasse, and R. Baratti. Measures of topological relevance based on the self-organizing map: Applications to process monitoring from spectroscopic measurements. In *EANN 2007, International Conference on Engineering Applications of Neural Networks, Thessaloniki (Greece)*, pages 24–33, 2007.

[14] F. Corona, E. Liitiäinen, A. Lendasse, and R. Baratti. Using functional representations in spectrophotoscopic variables selection and regression. In *SSC10, 10th Scandinavian Symposium on Chemometrics, Lappeenranta (Finland)*, page 29, 2007.

[15] F. Corona, L. Sassu, S. Melis, and R. Baratti. Measure of topological relevance for soft sensing product properties. In *DYCOPS 2007, IFAC International Symposium on Dynamics and Control of Process Systems, Cancun (Mexico)*, pages 175–180, 2007.

[16] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraclar, and A. Stolcke. Analysis of morph-based speech recognition and the modeling of out-of-vocabulary words across languages. In *Human Language Technologies / The Annual Conference of the North American Chapter of the Association for Computational Linguistics 2007*, pages 380–387, 2007.

[17] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraclar, and A. Stolcke. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Trans. Speech Lang. Process.*, 5(1 (Dec.)), 2007.

[18] M. Creutz and K. Lagus. Morfessor in the morpho challenge. In *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes 2006*, 2006.

[19] M. Creutz and K. Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):3, 2007.

[20] M. Creutz, K. Lagus, and S. Virpioja. Unsupervised morphology induction using morfessor. In A. Yli-Jyrä, L. Karttunen, and J. Karhumäki, editors, *Finite-State Methods and Natural Language Processing, Lecture Notes in Computer Science*, pages 300–301, Berlin, Heidelberg, 2006. Springer.

[21] T. Deselaers, A. Hanbury, V. Viitaniemi, A. Benczur, M. Brendel, B. Daroczy, E. Balderas, H. Jair, T. Gevers, H. Gracidas, C. Arturo, S. C. H. Hoi, J. Laaksonen, M. Li, M. Castro, H. Marisol, H. Ney, X. Rui, N. Sebe, J. Stöttinger, and L. Wu. Overview of the ImageCLEF 2007 object retrieval task. In *Working notes of the CLEF 2007 Workshop*, 2007.

[22] M. Everingham, A. Zisserman, C. K. I. Williams, L. V. Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkò, S. Duffner, J. Eichhorn, J. D. R. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, T. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. L. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. Storkey, S. Szedmak, B. T. Triggs, I. Ulusoy, V. Viitaniemi, and J. Zhang. The 2005 PASCAL visual object classes challenge. In Magnini and

F. d'Alche Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment*, pages 117–176, Berlin/Heidelberg, 2006. Springer.

[23] R. Girdziusas and J. Laaksonen. How marginal likelihood inference unifies entropy, correlation and SNR-based stopping in nonlinear diffusion scale-spaces. In I. S. Kweon, Y. Yagi, S. B. Kang, and H. E. Zha, editors, *Proc. of Asian Conf. on Computer Vision, Lecture Notes in Computer Science, Vol. 4843*, pages 811–820. Springer, 2007.

[24] R. Girdziusas and J. Laaksonen. When is a discrete diffusion a scale-space? In D. Metaxas, B. C. Vemuri, A. Shashua, and H. E. Shum, editors, *Proc. of 11th Int. Conf. on Computer Vision*. IEEE, 2007.

[25] A. Gross, S.-L. Joutsiniemi, R. Rimon, and B. Appelberg. Correlation of symptom clusters of schizophrenia with absolute powers of main frequency bands in quantitative EEG. *Behavioral and Brain Functions*, 2:23, 2006.

[26] R. Hakala, T. Similä, M. Sirola, and J. Parviainen. Process state and progress visualization using self-organizing map. In E. e. a. Corchado, editor, *IDEAL 2006, LNCS 4224*, pages 73–80, Berlin Heidelberg, 2006. Springer-Verlag.

[27] D. R. Hardoon, J. Shawe-Taylor, A. Ajanki, K. Puolamäki, and S. Kaski. Information retrieval by inferring implicit queries from eye movements. In M. Meila and X. Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

[28] M. Harva. A variational em approach to predicting uncertainty in supervised learning. In *World Congress on Computational Intelligence (WCCI'06)*, pages 11091–11095, 2006.

[29] M. Harva. A variational em approach to predictive uncertainty. *Neural Networks*, 20(4):550–558, 2007.

[30] M. Harva and A. Kaban. Variational learning for rectified factor analysis. *Signal Processing*, 87(3):509–527, 2007.

[31] M. Harva and S. Raychaudhury. Bayesian estimation of time delays between unevenly sampled signals. In *Int. Workshop on Machine Learning for Signal Processing (MLSP'06)*, pages 111–116, 2006.

[32] T. Hirsimäki. On compressing n-gram language models. In *IEEE International Conference on Acoustics 2007*, pages 949–952, 2007.

[33] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen. Unlimited vocabulary speech recognition with morph language models applied to finnish. *Computer, Speech and Language*, 20(4):515–541, 2006.

[34] J. Hollmén and J. Tikka. Compact and understandable descriptions of mixture of Bernoulli distributions. In M. Berthold, J. Shawe-Taylor, and N. Lavrac, editors, *Proceedings of the 7th International Symposium on Intelligent Data Analysis (IDA 2007)*, pages 1–12, Ljubljana, 2007. Springer-Verlag.

[35] A. Honkela. Distributed Bayes blocks for variational Bayesian learning. In *High Performance Computing for Statistical Inference 2006*, 2006.

[36] A. Honkela, J. Seppä, and E. Alhoniemi. Agglomerative independent variable group analysis. In *15th European Symposium on Artificial Neural Networks (ESANN 2007) 2007*, pages 55–60, 2007.

[37] A. Honkela, M. Tornio, T. Raiko, and J. Karhunen. Natural conjugate gradient in variational inference. Technical Report Report E10, Espoo, Finland, 2007.

[38] A. Honkela, H. Valpola, A. Ilin, and J. Karhunen. Blind separation of nonlinear mixtures by variational Bayesian learning. *Digital Signal Processing*, 17(5):914–934, 2007.

[39] T. Honkela. Models for self-organization of health care. In H. Toiviainen, F. Raitso, and K. Kosonen, editors, *Proceedings of FISCAR'07, abstracts*, page 80, Helsinki, 2007. University of Helsinki.

[40] T. Honkela. Philosophical aspects of neural, probabilistic and fuzzy modeling of language use and translation. In *Proceedings of IJCNN'07*. INNS, 2007.

[41] T. Honkela, V. Könönen, T. Lindh-Knuutila, and M.-S. Paukkeri. Simulating processes of language emergence, communication and agent modeling. In E. Hyvönen, T. Kauppinen, J. Kortela, M. Laukkanen, T. Raiko, and K. Viljanen, editors, *New Developments in Artificial Intelligence and the Semantic Web*, pages 129–132, Espoo, 2006. Finnish Artificial Intelligence Society.

[42] T. Honkela and M. Pöllä. Describing rich content: Future directions for the semantic web. In E. Hyvönen, T. Kauppinen, J. Kortela, M. Laukkanen, T. Raiko, and K. Viljanen, editors, *New Developments in Artificial Intelligence and the Semantic Web*, pages 143–148, Espoo, 2006. Finnish Artificial Intelligence Society.

[43] T. Honkela, T. Raiko, J. Kortela, and H. Valpola. *Proceedings of the Ninth Scandinavian Conference on Artificial Intelligence (SCAI 2006)*. Suomen tekoälyseura, Espoo, 2006.

[44] A. Ilin. Independent dynamics subspace analysis. In *14th European Symposium on Artificial Neural Networks 2006*, pages 345–350, 2006.

[45] A. Ilin, H. Valpola, and E. Oja. Exploratory analysis of climate data using source separation methods. *Neural Networks*, 19:155–167, 2006.

[46] A. Ilin, H. Valpola, and E. Oja. Extraction of climate components with structured variance. In *Proceedings of the IEEE World Congress on Computational Intelligence, WCCI 2006*, pages 10528–10535, 2006.

[47] A. Ilin, H. Valpola, and E. Oja. Finding interesting climate phenomena by exploratory statistical techniques. In *Fifth Conference on Artificial Intelligence Applications to Environmental Science 2007*, 2007.

[48] M. Inki. An easily computable eight times overcomplete ICA method for image data. In *Int. Conf. on Independent Component Analysis and Blind Signal Separation 2006*, 2006.

[49] N. Janasik and T. Honkela. Self-organizing map in qualitative research. In H. Toiviainen, F. Raitso, and K. Kosonen, editors, *Proceedings of FISCAR'07, abstracts*, page 79, Helsinki, 2007. University of Helsinki.

[50] J. Karhunen and T. Ukkonen. Generalizing independent component analysis for two related data sets. In *2006 IEEE World Congress on Computational Intelligence*. IEEE Computational Intelligence Society, 2006.

[51] J. Karhunen and T. Ukkonen. Extending ICA for finding jointly dependent components from two related data sets. *Neurocomputing*, 70(16-18):2969–2979, 2007.

[52] S. Kaski. Implicit feedback from eye movements for proactive information retrieval. In *NIPS 2006 Workshop on User Adaptive Systems 2006 http://www.ece.mcgill.ca/˜smanno1///UserAdaptiveSystems.html*, 2006.

[53] S. Kaski. Ydintehtävät kunniaan. *Polysteekki*, (3):19, 2006.

[54] S. Kaski and J. Peltonen. Learning from relevant tasks only. In J. N. Kok, J. Koronacki, R. L. de Mantaras, S. Matwin, D. Mladenic, and A. Skowron, editors, *Machine Learning: ECML 2007 (Proceedings of the 18th European Conference on Machine Learning)*, pages 608–615, Berlin, 2007. Springer-Verlag.

[55] S. Kaski, J. Rousu, and E. Ukkonen. Probabilistic modeling and machine learning in structural and systems biology. *BMC Bioinformatics*, 8 (Suppl2):1–2, 2007.

[56] K. Kersting, L. De Raedt, and T. Raiko. Logical hidden markov models. *Journal of Artificial Intelligence Research*, 25:425–456, 2006.

[57] H. Keski-Säntti, T. Atula, J. Tikka, J. Hollmén, A. A. Mäkitie, and I. Leivo. Predictive value of histopathologic parameters in early squamous cell carcinoma of oral tongue. *Oral Oncology*, 2006.

[58] K. Kettunen, M. Sadeniemi, T. Lindh-Knuutila, and T. Honkela. Analysis of eu languages through text compression. In T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala, editors, *Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006*, pages 99–109. Springer, 2006.

[59] A. Klami and S. Kaski. Generative models that discover dependencies between data sets. In *2006 IEEE International Workshop on Machine Learning for Signal Processing 2006*, pages 123–128. IEEE, 2006.

[60] A. Klami and S. Kaski. Generative models that discover dependencies between data sets. Technical Report Report E2, Helsinki University of Technology, Espoo, Finland, 2006.

[61] A. Klami and S. Kaski. Local dependent components. In Z. Ghahramani, editor, *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, pages 425–433. Omni Press, 2007.

[62] M. Klami and T. Honkela. Self-organized ordering of terms and documents in nsf awards data. In *Workshop on Self-Organizing Maps (WSOM 2007) 2007*, 2007.

[63] M. Klami and K. Lagus. Unsupervised word categorization using self-organizing maps and automatically extracted morphs. In *7th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2006) 2006*, pages 912–919, 2006.

[64] T. Knuuttila, A.-M. Rusanen, and T. Honkela. Self-organizing maps as travelling computational templates. In *Proceedings of IJCNN'07*, pages 1231–1236, 2007.

[65] T. Kohonen. Self-organizing neural projections. *Neural Networks*, 19:723–733, 2006.

[66] T. Kohonen. Description of input patterns by linear mixtures of SOM models. Technical Report Report E8, Espoo, Finland, 2007.

[67] T. Kohonen and T. Honkela. Kohonen network. In E. M. Izhikevich, editor, *Scholarpedia*, page 1568, San Diego, California, 2007. The Neurosciences Institute, San Diego, California.

[68] Z. Koldovský, P. Tichavský, and E. Oja. Efficient variant of algorithm FastICA for independent component analysis attaining the cramér-rao lower bound. *IEEE Transactions on Neural Networks*, 17(5):1265–1277, 2006.

[69] S. Kollias, A. Stafylopatis, W. Duch, and E. E. , Oja. *Artifical Neural Networks - ICANN 2006, Volume I, Lecture Notes in Computer Science, 4131.* Springer, Berlin, Germany, 2006.

[70] S. Kollias, A. Stafylopatis, W. Duch, and E. E. Oja. *Artifical Neural Networks - ICANN 2006, Volume II, Lecture Notes in Computer Science, 4132.* Springer, Berlin, Germany, 2006.

[71] M. Korpela and J. Hollmén. Extending an algorithm for clustering gene expression time series. In J. Rousu, S. Kaski, and E. Ukkonen, editors, *Probabilistic Modeling and Machine Learning in Structural and Systems Biology 2006*, pages 120–124, Helsinki, 2006.

[72] H. Kortejärvi, J. Malkki, M. Marvola, A. Urtti, M. Yliperttula, and P. Pajunen. Level a in vitro-in vivo correlation (ivivc) model with Bayesian approach to formulation series. *Journal of Pharmaceutical Sciences*, 95(7):1595–1605, 2006.

[73] M. Koskela and J. Laaksonen. Semantic concept detection from news videos with self-organizing maps. In *Proceedings of 3rd IFIP Conference on Artificial Intelligence Applications and Innovations*, pages 591–599, 2006.

[74] M. Koskela, M. Sjöberg, J. Laaksonen, V. Viitaniemi, and H. Muurinen. Rushes summarization with self-organizing maps. In *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, pages 45–49, Augsburg, 2007. ACM Press.

[75] M. Koskela, M. Sjöberg, V. Viitaniemi, J. Laaksonen, and P. Prentis. PicSOM experiments in trecvid 2007. In *Proceedings of the TRECVID 2007 Workshop*, Gaithersburg, 2007.

[76] M. Koskela and A. F. Smeaton. An empirical study of inter-concept similarities in multimedia ontologies. In *Proceedings of the 6th ACM international conference on Image and video retrieval (CIVR 2007)*, pages 464–471, Amsterdam, 2007. ACM Press.

[77] M. Koskela, A. F. Smeaton, and J. Laaksonen. Measuring concept similarities in multimedia ontologies: Analysis and evaluations. *IEEE Transactions on Multimedia*, 9(5):912–922, 2007.

[78] M. Kurimo, M. Creutz, and K. Lagus. *Proceedings of the PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes.* PASCAL European Network of Excellence, Venice, Italy, 2006.

[79] M. Kurimo, M. Creutz, and V. Turunen. Overview of morpho challenge in clef 2007. In A. Nardi and C. Peters, editors, *Working Notes for the CLEF 2007 Workshop*. CLEF, 2007.

[80] M. Kurimo, M. Creutz, and V. Turunen. Unsupervised morpheme analysis evaluation by ir experiments - morpho challenge 2007. In A. Nardi and C. Peters, editors, *Working Notes for the CLEF 2007 Workshop*. CLEF, 2007.

[81] M. Kurimo, M. Creutz, and M. Varjokallio. Unsupervised morpheme analysis evaluation by a comparison to a linguistic gold standard - morpho challenge 2007. In A. Nardi and C. Peters, editors, *Working Notes for the CLEF 2007 Workshop*. CLEF, 2007.

[82] M. Kurimo, M. Creutz, M. Varjokallio, E. Arisoy, and M. Saraclar. Unsupervised segmentation of words into morphemes - challenge 2005, an introduction and evaluation report. In *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes 2006*, 2006.

[83] M. Kurimo, M. Creutz, M. Varjokallio, E. Arisoy, and M. Saraclar. Unsupervised segmentation of words into morphemes - morpho challenge 2005: Application to automatic speech recognition. In *International Conference on Spoken Language Processing - Interspeech 2006 (ICSLP), Pittsburgh, Pennsylvania, USA, September 17-21*, 2006.

[84] M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar. Unlimited vocabulary speech recognition for agglutinative languages. In *Human Language Technology 2006*, 2006.

[85] T. Kärnä, F. Corona, and A. Lendasse. Compressing spectral data using optimized gaussian basis. In *Proceedings of Chimiometrie 2007, Lyon (France)*, 2007.

[86] T. Kärnä and A. Lendasse. Comparison of fda based time series prediction methods. In *ESTSP 2007, European Symposium on Time Series Prediction, Espoo (Finland)*, pages 77–86, 2007.

[87] T. Kärnä and A. Lendasse. Gaussian fitting based fda for chemometrics. In F. Sandoval, A. Prieto, J. Cabestany, M. Graña, editors, *Computational and Ambient Intelligence, 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 2007, Proceedings*, pages 186–193. Springer-Verlag, 2007.

[88] T. Kärnä and A. Lendasse. Optimal gaussian basis functions for chemometrics. In *SSC10, 10th Scandinavian Symposium on Chemometrics, Lappeenranta (Finland)*, page 79, 2007.

[89] T. Kärnä, F. Rossi, and A. Lendasse. LS-SVM functional network for time series prediction. In *ESANN 2006 2006*, pages 473–478, 2006.

[90] V. Könönen. Dynamic pricing based on asymmetric multiagent reinforcement learning. *International Journal of Intelligent Systems*, 21:73–98, 2006.

[91] J. Laaksonen, M. Koskela, M. Sjöberg, V. Viitaniemi, and H. Muurinen. Video summarization with SOMs. In *Proceedings of the 6th Int. Workshop on Self-Organizing Maps (WSOM 2007)*, Bielefeld, 2007.

[92] J. Laaksonen and V. Viitaniemi. Emergence of ontological relations from visual data with self-organizing maps. In *Proceedings of the 9th Scandinavian Conference on Artificial Intelligence (SCAI 2006)*, pages 31–38, 2006.

[93] K. Lagus and T. Honkela. Kognitiivinen mallintaminen. In M. Hämäläinen, M. Laine, O. Aaltonen, and A. Revonsuo, editors, *Mieli ja aivot. Kognitiivisen neurotieteen oppikirja*, pages 61–70, Turku, 2006. Kognitiivisen neurotieteen tutkimuskeskus, Turun yliopisto.

[94] D. T. H. Lai, J. Pakkanen, R. Begg, and M. Palaniswami. Computational intelligence and sensor networks: A convergence of technologies for future biomedical system. In N. Wickramasinghe, editor, *Encyclopedia of Health and Information Systems*, Australia, 2007. Idea House Publishing.

[95] A. Lendasse. *Proceedings of the European Symposium on Time Series Prediction (ESTSP'07)*. Espoo, 2007.

[96] A. Lendasse and F. Corona. Optimal linear projection based on noise variance estimation. In *Proceedings of Chimiometrie 2007, Lyon (France)*, 2007.

[97] A. Lendasse and F. Corona. Optimal linear projection based on noise variance estimation - application to spectrometric modeling. In *SSC10, 10th Scandinavian Symposium on Chemometrics, Lappeenranta (Finland)*, page 26, 2007.

[98] A. Lendasse, F. Corona, J. Hao, N. Reyhani, and M. Verleysen. Determination of the mahalanobis matrix using nonparametric noise estimations. In *ESANN 2006 2006*, pages 227–232, 2006.

[99] A. Lendasse, F. Corona, S.-P. Reinikainen, and P. Minkkinen. Functional variable selection using noise variance estimation. In *Proceedings of Chimiometrie 2007, Lyon (France)*, 2007.

[100] A. Lendasse and E. Liitiäinen. Variable scaling for time series prediction: Application to the estsp'07 and the nn3 forecasting competitions. In *IJCNN 2007, International Joint Conference on Neural Networks, Orlando, Florida, USA*, pages 2812–2816, 2007.

[101] A. Lendasse, E. Oja, O. Simula, and M. Verleysen. Time series prediction competition: The cats benchmark. *Neurocomputing*, 70(13-15):2325–2329, 2007.

[102] E. Liitiäinen, F. Corona, and A. Lendasse. Nearest neighbor distributions and noise variance estimation. In *ESANN 2007, European Symposium on Artificial Neural Networks, Bruges (Belgium)*, pages 67–72, 2007.

[103] E. Liitiäinen, F. Corona, and A. Lendasse. Non-parametric residual variance estimation in supervised learning. In F. Sandoval, A. Prieto, J. Cabestany, M. Graña, editors, *Computational and Ambient Intelligence, 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 2007, Proceedings*, pages 63–71. Springer-Verlag, 2007.

[104] E. Liitiäinen and A. Lendasse. Long-term prediction of time series using state-space models. In *Lecture Notes in Computer Science, Volume 4132/2006*, pages 181–190. Springer Berlin / Heidelberg, 2006.

[105] E. Liitiäinen, N. Reyhani, and A. Lendasse. Em-algorithm for training of state-space models with application to time series prediction. In *ESANN 2006 2006*, pages 137–142, 2006.

[106] T. Lindh-Knuutila, T. Honkela, and K. Lagus. Simulating meaning negotiation using observational language games. In P. e. a. Vogt, editor, *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication*, pages 168–179, Berlin/Heidelberg, 2006. Springer.

[107] P. Lindholm, P. Nymark, H. Wikman, K. Salmenkivi, A. Nicholson, M. V. Korpela, S. Kaski, S. Ruosaari, J. Hollmén, E. Vanhala, A. Karjalainen, S. Anttila, V. Kinnula, and S. Knuutila. Asbestos-associated malignancies in the lung and pleura show distinct genetic aberrations. *Lung cancer*, 54:15, 2006.

[108] L. Lindqvist, T. Honkela, and M. Pantzar. Visualizing practice theory through a simulation model. Technical Report Report E9, Helsinki University of Technology, Computer and Information Science, Espoo, 2007.

[109] S. Luyssaert, I. Janssens, M. Sulkava, D. Papale, A. Dolman, M. Reichstein, T. Suni, J. Hollmén, T. Vesala, D. Lousteau, B. Law, and E. Moors. Photosynthesis drives interannual variability in net carbon-exchange of pine forests at different latitudes. In *Proceedings of the Open Science Conference on the GHG Cycle in the Northern Hemisphere*, pages 86–87, Jena, Germany, 2006. CarboEurope, NitroEurope, CarboOcean, and Global Carbon Project.

[110] S. Luyssaert, I. A. Janssens, M. Sulkava, D. Papale, A. J. Dolman, M. Reichstein, J. Hollmén, J. G. Martin, T. Suni, T. Vesala, D. Lousteau, B. E. Law, and E. J. Moors. Photosynthesis drives anomalies in net carbon-exchange of pine forests at different latitudes. *Global Change Biology*, 13(10):2110–2127, 2007.

[111] S. Luyssaert, M. Sulkava, H. Raitio, J. Hollmén, and P. Merilä. Is n and s deposition altering the mineral nutrient composition of norway spruce and scots pine needles in finland? In J. Eichhorn, editor, *Proceedings of Symposium: Forests in a Changing Environment - Results of 20 years ICP Forests Monitoring*, pages 80–81, Göttingen, Germany, 2006. ICP Forests, European Commission, Nordwestdeutsche Versuchsanstalt.

[112] N. Matsuda, J. Laaksonen, F. Tajima, N. Miyatake, and H. Sato. Comparison with observer appraisals of fundus images and diagnosis by using learning vector quantization. In *Proceedings of the 23rd Fuzzy System Symposium*, Nagoya, 2007.

[113] M. A. Mayer, V. Karkaletsis, K. Stamatakis, A. Leis, D. Villarroel, C. Thomeczek, M. Labsky, F. , Ló pez-Ostenero, and T. Honkela. Medieq - quality labelling of medical web content using multilingual information extraction. In L. Bos, L. Roa, K. Yogesan, B. O'Connell, A. Marsh, and B. Blobel, editors, *Medical and Care Compunetics 3*, pages 183–190, Amsterdam, Netherlands, 2006. IOS Press.

[114] Y. Miche, P. Bas, A. Lendasse, C. Jutten, and O. Simula. Advantages of using feature selection techniques on steganalysis schemes. In F. Sandoval, A. Prieto, J. Cabestany, M. Graña, editors, *Computational and Ambient Intelligence, 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 2007, Proceedings*, pages 606–613. Springer-Verlag, 2007.

[115] Y. Miche, P. Bas, A. Lendasse, O. Simula, and C. Jutten. Avantages de la sélection de caractéristiques pour la stéganalyse. In *GRETSI 2007, Groupe de Recherche et d'Etudes du Traitement du Signal et des Images, Troyes, France*, 2007.

[116] Y. Miche, B. Roue, A. Lendasse, and P. Bas. A feature selection methodology for steganalysis. In *Lecture Notes in Computer Science*, pages 49–59, Berlin / Heidelberg, Istanbul, 2006. Springer.

[117] M. Molinier, J. Laaksonen, and T. Häme. Self-organising maps for change detection and monitoring of human activity in satellite imagery. In *Proceedings of ESA-EUSC 2006*, 2006.

[118] M. Molinier, J. Laaksonen, and T. Häme. A self-organizing map framework for detection of man-made structures and changes in satellite imagery. In *Proceedings of IEEE International Geoscience And Remote Sensing Symposium*, 2006.

[119] M. Molinier, J. Laaksonen, and T. Häme. Detecting man-made structures and changes in satellite imagery with a content-based information retrieval system built on self-organizing maps. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4):861–874, 2007.

[120] M. Molinier, J. Laaksonen, Y. Rauste, and T. Häme. Detecting changes in polarimetric sar data with content-based image retrieval. In *Proceedings of IEEE International Geoscience And Remote Sensing Symposium*, Barcelona, 2007.

[121] M. Multanen, K. Raivio, and P. Lehtimäki. Hierarchical analysis of GSM network performance data. In *ESANN 2006*, pages 449–454, 2006.

[122] H. Muurinen and J. Laaksonen. Video segmentation and shot boundary detection using self-organizing maps. In *Proceedings of 15th Scandinavian Conference on Image Analysis (SCIA 2007)*, pages 770–779, Aalborg, 2007.

[123] J. Nikkilä, A. Honkela, and S. Kaski. Exploring the independence of gene regulatory modules. In J. Rousu, S. Kaski, and E. Ukkonen, editors, *Proceedings of Probabilistic Modeling and Machine Learning in Structural and Systems Biology (PMSB 2006)*, pages 131–136. Helsinki University Printing House, 2006.

[124] L. Nolan, M. Harva, A. Kabán, and S. Raychaudhury. A data-driven Bayesian approach for finding young stellar populations in early-type galaxies from their uv-optical spectra. *Monthly Notices of the Royal Astronomical Society*, 366(1):321–338, 2006.

[125] K. Nybo, J. Venna, and S. Kaski. The self-organizing map as a visual neighbor retrieval method. In *6th International Workshop on Self-Organizing Maps 2007*, 2007.

[126] P. Nymark, P. M. Lindholm, M. V. Korpela, L. Lahti, S. Ruosaari, S. Kaski, J. Hollmén, S. Anttila, V. L. Kinnula, and S. Knuutila. Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines. *BMC Genomics*, 8:1–14, 2007.

[127] E. Oja, M. Sjöberg, V. Viitaniemi, and J. Laaksonen. Emergence of semantics from multimedia databases. In *Computational Intelligence: Principles and Practice*. IEEE Computational Intelligence Society, 2006.

[128] E. Oja and Z. Yuan. The fastica algorithm revisited - convergence analysis. *IEEE Transactions on Neural Networks*, 17(6):1370–1381, 2006.

[129] M. Oja. In silico expression profiles of human endogenous retroviruses. In J. Ragapakse, B. Schmidt, and G. E. Volkert, editors, *Proc. of Workshop on Pattern Recognition in Bioinformatics (PRIB 2007), Lecture Notes in Bioinformatics, Vol. 4774*, pages 253–263. Springer, 2007.

[130] M. Oja, J. Peltonen, J. Blomberg, and S. Kaski. Methods for estimating human endogenous retrovirus activities from est databases. *BMC Bioinformatics*, 8 (Suppl2):1–12, 2007.

[131] M. Oja, J. Peltonen, and S. Kaski. Estimation of human endogenous retrovirus activities from expressed sequence databases. In J. Rousu, S. Kaski, and E. Ukkonen, editors, *Probabilistic Modeling and Machine Learning in Structural and Systems Biology (PMSB 2006) 2006*, pages 50–54, Helsinki, Finland, 2006. University of Helsinki.

[132] J. Pakkanen, J. Iivarinen, and E. Oja. The evolving tree - analysis and applications. *IEEE Transactions on Neural Networks*, 17(3):591–603, 2006.

[133] J. Pakkanen, D. T. H. Lai, and M. Palaniswami. A study on multidimensional scaling (mds) algorithms for 3d sensor network localisation. In *ISSNIP/ISPRS Joint International Workshop on Distributed Geoinformatics and Sensing 2007*, 2007.

[134] J. Pakkanen and T. H. Lai, Daniel. On the transmission complexity of centralized and distributed localisation schemes. In *Third International Conference on Intelligent Sensor Networks and Information Processing 2007*, 2007.

[135] K. J. Palomäki, G. J. Brown, and J. Barker. Recognition of reverberant speech using full cepstral features and spectral missing data. In *IEEE International Conference on Acoustics 2006*, pages 289–292, 2006.

[136] M. Pantzar, T. Honkela, I. Nieminen, and A. Wallenius. Dynamic visualization of practice theory. In H. Toiviainen, F. Raitso, and K. Kosonen, editors, *Proceedings of FISCAR'07, abstracts*, page 77, Helsinki, 2007. University of Helsinki.

[137] J. Peltonen, J. Goldberger, and S. Kaski. Fast discriminative component analysis for comparing examples. In *NIPS 2006 workshop on Learning to Compare Examples 2006*, 2006.

[138] J. Peltonen, J. Goldberger, and S. Kaski. Fast semi-supervised discriminative component analysis. In K. Diamantaras, T. Adali, I. Pitas, J. Larsen, T. Papadimitriou, and S. Douglas, editors, *Machine Learning for Signal Processing XVII*, pages 312–317. IEEE, 2007.

[139] J. Peltonen and S. Kaski. Learning when only some of the training data are from the same distribution as test data. In *NIPS 2006 Workshop on Learning when Test and Training Inputs Have Different Distributions 2006*, 2006.

[140] M. Pesonen, M. Laine, R. Vigário, and C. Krause. Brain oscillatory EEG responses reflect auditory memory functions. In *13th World Congress of Psychophysiology 2006*, 2006.

[141] A. Pietilä, M. El-Segaier, R. Vigário, and E. Pesonen. Blind source separation of cardiac murmurs from heart recordings. In *6th Int. Conf. on Independent Component Analysis and Blind Signal Separation Charleston 2006*, pages 470–477, 2006.

[142] K. Puolamäki and S. Kaski. *Proceedings of the NIPS 2005 Workshop on Machine Learning for Implicit Feedback and User Modeling*. Espoo, 2006.

[143] K. Puolamäki, J. Salojärvi, E. Savia, and S. Kaski. Discriminative mcmc. Technical Report Report E1, Helsinki University of Technology, Espoo, Finland, 2006.

[144] A. Puurula and M. Kurimo. Vocabulary decomposition for estonian open vocabulary speech recognition. In *45th Annual Meeting of the Association for Computational Linguistics 2007*. ACL, 2007.

[145] J. Pylkkönen. Lda based feature estimation methods for lvcsr. In *9th International Conference on Spoken Language Processing (Interspeech 2006) 2006*, pages 389–392, 2006.

[146] J. Pylkkönen. Estimating vtln warping factors by distribution matching. In *8th Annual Conference of the International Speech Communication Association (Interspeech 2007) 2007*, pages 270–273, 2007.

[147] M. Pöllä and T. Honkela. Modeling anticipatory behavior with self-organizing neural networks. In D. Dubois, editor, *Computing Anticipatory Systems (CASYS'05)*, Woodbury, New York, 2006. American Institute of Physics.

[148] M. Pöllä and T. Honkela. Probabilistic text change detection using an immune model. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2007*, pages 1109–1114, 2007.

[149] M. Pöllä, T. Honkela, H. Bruun, and A. Russell. Analysis of interdisciplinary text corpora. In *Proceedings of Ninth Scandinavian Conference on Artificial Intelligence (SCAI 2006)*, pages 17–22, Espoo, Finland, 2006.

[150] T. Raiko. Higher order statistics in play-out analysis. In T. Honkela, T. Raiko, J. Kortela, and H. Valpola, editors, *Proceedings of the Scandinavian Conference on Artificial Intelligence, SCAI 2006*, pages 189–195, Espoo, 2006. Suomen tekoälyseura.

[151] T. Raiko, A. Ilin, and J. Karhunen. Principal component analysis for large scale problems with lots of missing values. In J. e. a. Kok, editor, *Lecture Notes in Artificial Intelligence, vol. 4701, Springer-Verlag, proceedings of the 18th European Conference on Machine Learning (ECML 2007)*, pages 691–698. Springer-Verlag, 2007.

[152] T. Raiko, M. Tornio, A. Honkela, and J. Karhunen. State inference in variational Bayesian nonlinear state-space models. In D. E. Justinian Rosca and S. H. José C. Principe, editors, *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Source Separation*, pages 222–229. Springer, 2006.

[153] T. Raiko, H. Valpola, M. Harva, and J. Karhunen. Building blocks for variational Bayesian learning of latent variable models. Technical Report Report E4, Helsinki University of Technology, Espoo, 2006.

[154] T. Raiko, H. Valpola, M. Harva, and J. Karhunen. Building blocks for variational Bayesian learning of latent variable models. *Journal of Machine Learning Research*, 8:155–201, 2007.

[155] K. Raivio. Analysis of soft handover measurements in 3G network. In E. Alba, C.-F. Chiasserini, N. Abu-Ghazaleh, and R. L. Cigno, editors, *Proceedings of the Ninth ACM Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 330–337, New York, 2006. ACM Press.

[156] K. Raju, T. Ristaniemi, J. Karhunen, and E. Oja. Jammer suppression in ds-cdma arrays using independent component analysis. *IEEE Transactions on Wireless Communications*, 5(1):77–82, 2006.

[157] U. Remes, J. Pylkkönen, and M. Kurimo. Segregation of speakers for speaker adaptation in tv news audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.*, pages 481–484. IEEE, 2007.

[158] N. Reyhani and A. Lendasse. An empirical dependence measures based on residual variance estimation. In *ISSPA 2007, International Symposium on Signal Processing and its Applications in conjunction with the International Conference on Information Sciences, Signal Processing and its Applications, Sharjah, United Arab Emirates (U.A.E.)*, 2007.

[159] R. Ritala, E. Alhoniemi, T. Kauranne, K. Konkarikoski, A. Lendasse, and M. Sirola. Nonlinear temporal and spatial forecasting: modeling and uncertainty analysis (notes) - masit20. In E. Alakangas and P. Taskinen, editors, *MASI Technology Programme 2005 - 2009 Yearbook 2007*, page 188, Helsinki, 2007. Tekes.

[160] F. Rossi, A. Lendasse, François D., V. Wertz, and M. Verleysen. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and Intelligent Laboratory Systems*, 80(2):215–226, 2006.

[161] J. Rousu, S. Kaski, and E. Ukkonen. *Probabilistic Modeling and Machine Learning in Structural and Systems Biology.* University of Helsinki, Helsinki, 2006.

[162] J.-H. Schleimer and R. Vigário. Reference-based extraction of phase synchronous components. In *16th Int. Conf. on Artificial Neural Networks (ICANN'2006)*, pages 230–238, 2006.

[163] U. Seiffert, B. Hammer, S. Kaski, and T. Villmann. Neural networks and machine learning in bioinformatics - theory and applications. In *Proceedings of ESANN'06, 14th European Symposium on Artificial Neural Networks*, pages 521–532, Evere, Belgium, 2006. d-side.

[164] V. Siivola, M. Creutz, and M. Kurimo. Morfessor and varikn machine learning tools for speech and language technology. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association*, pages 1549–1552. ISCA, 2007.

[165] V. Siivola, T. Hirsimäki, and S. Virpioja. On growing and pruning kneser-ney smoothed n-gram models. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5):1617–1624, 2007.

[166] T. Similä. Self-organizing map visualizing conditional quantile functions with multidimensional covariates. *Computational Statistics & Data Analysis*, 50(8):2097–2110, 2006.

[167] T. Similä. Majorize-minimize algorithm for multiresponse sparse regression. In *IEEE International Conference on Acoustics, Speech, and Signal Processing 2007*, pages 553–556, 2007.

[168] T. Similä and J. Tikka. Common subset selection of inputs in multiresponse regression. In *IEEE International Joint Conference on Neural Networks (IJCNN) 2006*, pages 3574–3581, 2006.

[169] T. Similä and J. Tikka. Input selection and shrinkage in multiresponse linear regression. Technical Report Report A85, Helsinki University of Technology, Espoo, 2006.

[170] T. Similä and J. Tikka. Input selection and shrinkage in multiresponse linear regression. *Computational Statistics & Data Analysis*, 52(1):406–422, 2007.

[171] O. Simula, A. Lendasse, F. Corona, S.-P. Reinikainen, M.-L. Riekkola, K. Hartonen, I. Vuorinen, and J. Silén. Developing chemometrics with the tools of information sciences (chess) - masit23. In E. Alakangas and P. Taskinen, editors, *MASI Technology Programme 2005-2009, Yearbook 2007*, pages 201–221, 2007.

[172] J. Sinkkonen, J. Aukia, and S. Kaski. Inferring vertex properties from topology in large networks. In *MLG'07, the 5th International Workshop on Mining and Learning with Graphs, Firenze, Aug 1-3, 2007*, 2007.

[173] M. Sirola. Applying decision analysis method in accident management process control problem. *Systems science*, 32(1):122, 2006.

[174] M. Sirola. New methodologies and visualization techniques for npp control room concept. In *Workshop on Human System Interfaces - Design and Evaluation (HSI) 2006*. OECD Halden Reactor Project, 2006.

[175] M. Sirola, G. Lampi, and J. Parviainen. Failure detection and separation in SOM based decision support. In *Workshop on Self-Organizing Maps 2007*, 2007.

[176] M. Sjöberg, J. Laaksonen, M. Pöllä, and T. Honkela. Retrieval of multimedia objects by combining semantic information from visual and textual descriptors. In *Proceedings of 16th International Conference on Artificial Neural Networks*, pages 75–83, 2006.

[177] M. Sjöberg, H. Muurinen, J. Laaksonen, and M. Koskela. PicSOM experiments in trecvid 2006. In *Proceedings of the TRECVID 2006 Workshop*, pages 262–270, 2006.

[178] M. Sjöberg, V. Viitaniemi, J. Laaksonen, and T. Honkela. Analysis of semantic information available in an image collection augmented with auxiliary data. In I. Maglogiannis, K. Karpouzis, and M. Bramer, editors, *Proceedings of 3rd IFIP Conference on Artificial Intelligence Applications and Innovations*, pages 600–608, Athens, 2006. IFIP.

[179] P. Somervuo, A. Härmä, and S. Fagerlund. Parametric representations of bird sounds for automatic species recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6):2252–2263, 2006.

[180] A. Sorjamaa, J. Hao, N. Reyhani, Y. Ji, and A. Lendasse. Methodology for long-term prediction of time series. *Neurocomputing*, 70(16-18):2861–2869, 2007.

[181] A. Sorjamaa and A. Lendasse. Time series prediction using dirrec strategy. In *ESANN 2006 2006*, pages 143–148, 2006.

[182] A. Sorjamaa and A. Lendasse. Time series prediction as a problem of missing values. In *ESTSP 2007, European Symposium on Time Series Prediction, Espoo (Finland)*, pages 165–174, 2007.

[183] A. Sorjamaa, E. Liitiäinen, and A. Lendasse. Time series prediction as a problem of missing values: Application to estsp2007 and nn3 competition benchmarks. In *IJCNN 2007, International Joint Conference on Neural Networks, Orlando, Florida, USA*, pages 2948–2953, 2007.

[184] A. Sorjamaa, P. Merlin, B. Maillet, and A. Lendasse. A nonlinear approach for the determination of missing values in temporal databases. In *Mashs 2007, Computational Methods for Modelling and Learning in Social and Human Sciences, Brest (France)*, 2007.

[185] A. Sorjamaa, P. Merlin, B. Maillet, and A. Lendasse. SOM+eof for finding missing values. In *ESANN 2007, European Symposium on Artificial Neural Networks, Bruges (Belgium)*, pages 115–120, 2007.

[186] K. Stamatakis, K. Chandrinos, V. Karkaletsis, M. A. Mayer, D. V. Gonzales, M. Labsky, E. Amigo, and M. Pöllä. Aqua, a system assisting labelling experts assess health web resources. In *Proceedings of 12th International Symposium for Health Information Management Research, iSHIMR 2007*, 2007.

[187] M. Sulkava. Modeling how varying data quality affects the ability to detect trends in environmental time series. In V. Mäkinen, G. Lindén, and H. Toivonen, editors, *Summer School on Algorithmic Data Analysis (SADA 2007) and Annual Hecse Poster Session*, page 104, Helsinki, 2007. Helsinki University Printing House.

[188] M. Sulkava, S. Luyssaert, P. Rautio, I. A. Janssens, and J. Hollmén. Modeling the effects of varying data quality on trend detection in environmental monitoring. *Ecological Informatics*, 2(2):167–176, 2007.

[189] M. Sulkava, H. Mäkinen, P. Nöjd, and J. Hollmén. CUSUM charts for detecting onset and cessation of xylem formation based on automated dendrometer data. In I. Horová and J. Hrebícek, editors, *TIES 2007 - 18th annual meeting of the International Environmetrics Society*, page 111, Mikulov, 2007. Masaryk University.

[190] M. Sulkava, J. Tikka, and J. Hollmén. Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. *Ecological Modelling*, 191:118–130, 2006.

[191] P. Tichavský, Z. Koldovský, and E. Oja. Performance analysis of the FastICA for algorithm and cramér-rao bounds for linear independent component analysis. *IEEE Transactions on Signal Processing*, 54(4):1189–1203, 2006.

[192] P. Tichavský, Z. Koldovský, and E. Oja. Speed and accuracy enhancement of linear ICA techniques using rational nonlinear functions. In M. Davies, C. James, S. Abdallah, and M. Plumbley, editors, *Lecture notes in computer science vol. 4666*, pages 285–292, Berliini, 2007. Springer Verlag.

[193] J. Tikka. Input selection for radial basis function networks by constrained optimization. In J. M. de Sá, , L. A. Alexandre, W. Duch, and D. Mandic, editors, *Proceedings of the 17th International Conference on Artificial Neural Networks (ICANN 2007)*, pages 239–248, Porto, 2007.

[194] J. Tikka and J. Hollmén. Long-term prediction of time series using a parsimonious set of inputs and LS-SVM. In A. Lendasse, editor, *Proceedings of the First European Symposium on Time Series Prediction (ESTSP 2007)*, pages 87–96, Espoo, 2007.

[195] J. Tikka and J. Hollmén. A sequential input selection algorithm for long-term prediction of time series. *Neurocomputing*, 2007.

[196] J. Tikka, J. Hollmén, and S. Myllykangas. Mixture modeling of DNA copy number amplification patterns in cancer. In F. Sandoval, A. Prieto, J. Cabestany, and M. Grana, editors, *Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN 2007)*, pages 972–979, San Sebastián, 2007.

[197] J. Tikka, A. Lendasse, and J. Hollmén. Analysis of fast input selection: Application in time series prediction. In *International Conference on Artificial Neural Networks (ICANN) 2006*, pages 161–170, 2006.

[198] M. Tornio, A. Honkela, and J. Karhunen. Time series prediction with variational Bayesian nonlinear state-space models. In *European Symposium on Time Series Predictions (ESTSP 2007) 2007*, pages 11–19, 2007.

[199] M. Tornio and T. Raiko. Variational Bayesian approach for nonlinear identification and control. In M. Alamir and F. Allgöwer, editors, *Proceedings of the IFAC Workshop on Nonlinear Model Predictive Control for Fast Systems, NMPC FS06*, pages 41–46, Grenoble, 2006. Le Laboratoire d'Automatique de Grenoble.

[200] V. T. Turunen and M. Kurimo. Using latent semantic indexing for morph-based spoken document retrieval. In *9th International Conference on Spoken Language Processing (Interspeech 2006) 2006*, pages 341–344, 2006.

[201] V. T. Turunen and M. Kurimo. Indexing confusion networks for morph-based spoken document retrieval. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 631–638, New York, NY, USA, 2007. ACM Press.

[202] H. Valpola and A. Honkela. Hyperparameter adaptation in variational Bayes for the gamma distribution. Technical Report Report E6, Helsinki University of Technology, Espoo, 2006.

[203] J. Vandewalle, J. Suykens, B. De Moor, and A. Lendasse. State-of-the-art and evolution in public data sets and competitions for system identification, time series prediction and pattern recognition. In *32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hawaii Convention Center in Honolulu (USA)*, pages 1269–1272, 2007.

[204] M. Varjokallio and M. Kurimo. Comparison of subspace methods for gaussian mixture models in speech recognition. In *Proceedings of the INTERSPEECH 2007*, pages 2121–2124, 2007.

[205] J. Venna and S. Kaski. Local multidimensional scaling. *Neural Networks*, 19:889–899, 2006.

[206] J. Venna and S. Kaski. Nonlinear dimensionality reduction as information retrieval. Technical Report Report E5, Helsinki University of Technology, Espoo, Finland, 2006.

[207] J. Venna and S. Kaski. Visualizing gene interaction graphs with local multidimensional scaling. In *14th European Symposium on Artificial Neural Networks (ESANN'2006)*, pages 557–562, 2006.

[208] J. Venna and S. Kaski. Comparison of visualization methods for an atlas of gene expression data sets. *Information Visualization*, 6(2):139–154, 2007.

[209] J. Venna and S. Kaski. Nonlinear dimensionality reduction as information retrieval. In M. Meila and X. Shen, editors, *Proceedings of AISTATS 2007, the 11th International Conference on International Conference on Artificial Intelligence and Statistics*. Omnipress, 2007.

[210] V. Viitaniemi and J. Laaksonen. Focusing keywords to automatically extracted image segments using self-organising maps. In *Soft Computing in Image Processing: Recent Advances*, pages 121–156, Berliini/Heidelberg, 2006. Springer.

[211] V. Viitaniemi and J. Laaksonen. Techniques for still image scene classification and object detection. In *16th International Conference on Artificial Neural Networks (ICANN 2006) 2006*, pages 35–44. Springer, 2006.

[212] V. Viitaniemi and J. Laaksonen. Use of image regions in context-adaptive image classification. In *1st International Conference on Semantics And digital Media Technologies (SAMT 2006) 2006*, 2006.

[213] V. Viitaniemi and J. Laaksonen. Empirical investigations on benchmark tasks for automatic image annotation. In *Proceedings of the 9th International Conference on Visual Information Systems (VISUAL 2007)*, pages 96–107. Springer, 2007.

[214] V. Viitaniemi and J. Laaksonen. Evaluating performance of automatic image annotation: example case by fusing global image features. In *Proceedings of Fifth International Workshop on Content-Based Multimedia Indexing (CBMI 2007)*, pages 251–258, Bordeaux, 2007.

[215] V. Viitaniemi and J. Laaksonen. Evaluating the performance in automatic image annotation: Example case by adaptive fusion of global image features. *Signal Processing: Image Communication*, 22(6):557–568, 2007.

[216] V. Viitaniemi and J. Laaksonen. Improving the accuracy of global feature fusion based image categorisation. In *Proceedings of the 2nd International Conference on Semantic and Digital Media Technologies (SAMT 2007)*, pages 1–14, Genova, 2007. Springer.

[217] V. Viitaniemi and J. Laaksonen. Thoughts on evaluation of image retrieval inspired by imageclef 2007 object retrieval task. In *Proceedings of 3rd MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*, 2007.

[218] S. Virpioja and M. Kurimo. Compact n-gram models by incremental growing and clustering of histories. In *International Conference on Spoken Language Processing (Interspeech - ICSLP) 2006*, pages 1037–1040, 2006.

[219] S. Virpioja, J. J. Väyrynen, M. Creutz, and M. Sadeniemi. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Machine Translation Summit XI 2007*, pages 491–498, 2007.

[220] J. J. Väyrynen, T. Honkela, and L. Lindqvist. Towards explicit semantic features using independent component analysis. In M. Sahlgren and O. E. Knuttson, editors, *Proc. of the Workshop Semantic Content Acquisition and Representation, SCAR 2007, SICS Technical Report T2007-06*, Stockholm, Sweden, 2007. Swedish Institute of Computer Science.

[221] J. J. Väyrynen and T. Lindh-Knuutila. Emergence of multilingual representations by independent component analysis using parallel corpora. In T. Honkela, T. Raiko, J. Kortela, and H. Valpola, editors, *Proceedings of the Ninth Scandinavian Conference on Artificial Intelligence (SCAI 2006)*, pages 101–105, Espoo, 2006. Finnish Artificial Intelligence Society.

[222] J. J. Väyrynen, L. Lindqvist, and T. Honkela. Sparse distributed representations for words with thresholded independent component analysis. In *Proc. of International Joint Conference on Neural Networks (IJCNN 2007)*, pages 1031–1036, 2007.

[223] Z. Yang and J. Laaksonen. A fast fixed-point algorithm for two-class discriminative feature extraction. In *Proceedings of 16th International Conference on Artificial Neural Networks (ICANN 2006)*, pages 330–339, 2006.

[224] Z. Yang and J. Laaksonen. Approximated geodesic updates with principal natural gradients. In *Proceedings of The 2007 International Joint Conference on Neural Networks (IJCNN 2007)*, pages 1320–1325, Orlando, 2007.

[225] Z. Yang and J. Laaksonen. Face recognition using parzenfaces. In *Proceedings of International Conference on Artificial Neural Networks (ICANN'07)*, pages 200–209, Porto, 2007. Springer.

[226] Z. Yang and J. Laaksonen. Multiplicative updates for non-negative projections. *Neurocomputing*, 71(1-3):363–373, 2007.

[227] Z. Yang and J. Laaksonen. Regularized neighborhood component analysis. In *Proceedings of 15th Scandinavian Conference on Image Analysis (SCIA)*, pages 253–262, Aalborg, 2007.

[228] Z. Yang, Z. Yuan, and J. Laaksonen. Projective non-negative matrix factorization with applications to facial image processing. *International Journal of Pattern Recognition and Artificial Intelligence*, 21(8):1353–1362, 2007.

[229] J. Ylipaavalniemi, S. Mattila, A. Tarkiainen, and R. Vigário. Brains and phantoms: An ICA study of fmri. In *6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2006) 2006*, pages 503–510, 2006.

[230] J. Ylipaavalniemi, E. Savia, R. Vigário, and S. Kaski. Functional elements and networks in fmri. Technical Report Report E7, Helsinki University of Technology, Espoo, Finland, 2006.

[231] J. Ylipaavalniemi, E. Savia, R. Vigário, and S. Kaski. Functional elements and networks in fMRI. In *Proceedings of the 15th European Symposium on Artificial Neural Networks (ESANN 2007)*, pages 561–566, Bruges, 2007.

[232] J. Ylipaavalniemi and R. Vigário. Subspaces of spatially varying independent components in fMRI. In *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation (ICA 2007)*, pages 665–672, London, 2007.

[233] Q. Yu, E. Severin, and A. Lendasse. A global methodology for variable selection: Application to financial modeling. In *Mashs 2007, Computational Methods for Modelling and learning in Social and Human Sciences, Brest (France)*, 2007.

[234] Q. Yu, E. Severin, and A. Lendasse. Variable selection for financial modeling. In *CEF 2007, 13th International Conference on Computing in Economics and Finance Montréal, Quebec, Canada*, pages 115–126, 2007.

II–From Data to Knowledge Research Unit
Research Projects under the CIS Laboratory

Chapter 17

From Data to Knowledge Research Unit

Heikki Mannila, Jaakko Hollmén, Kai Puolamäki, Gemma Garriga, Jouni Seppänen, Robert Gwadera, Sami Hanhijärvi, Hannes Heikinheimo, Samuel Myllykangas, Antti Ukkonen, Nikolaj Tatti, Jarkko Tikka

17.1 Finding and using patterns

Finding trees

Hannes Heikinheimo, Jouni Seppänen, Nikolaj Tatti, Heikki Mannila

One of the most active topics in mining of binary data is finding interesting itemsets. A traditional approach is to search for frequent itemsets. In such sets the items co-occur frequently. While this definition of importance has many nice theoretical and practical properties it has one serious drawback: A frequent itemset, say AB, is less interesting if the individual attributes A and B are frequent. On the other hand, if AB is infrequent and A and B are frequent, then the itemset AB should be interesting.

One commonly used approach for defining the importance of the itemset is to compare the frequency against the independence assumption. The more the itemset deviates from the independence assumption, the more interesting it is. An alternative way of looking at this approach is to think that we are predicting the frequency an itemset from its individual attributes. In [1,2] we expand this idea by using the itemsets for the prediction. That is, we compare the itemset against a prediction based on a given family of itemsets. For prediction we use Maximum Entropy, a popular method for estimating distributions. For the comparison we use Kullback-Leibler divergence. We point out that this ranking method should be normalized and the normalization can be used as a statistical test.

We tested our methods with several real-world datasets. In our experiments we found out that a surprisingly large portion of itemsets that are important according to the independence model becomes unimportant when we are using larger itemsets for the prediction. However, in some cases using less itemsets produces a better ranking. This interesting phenomenon is a type of overlearning that occurs when too many itemsets mislead the prediction and model the noise in data (see Figure 17.1). To remedy this behavior we suggest in [1] a greedy algorithm that automatically picks the lowest rank for the itemset.

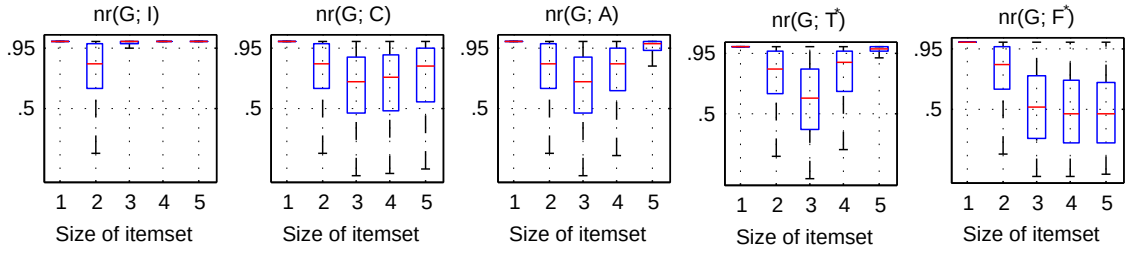


Figure 17.1: Box plots of the rank measures computed from *Paleo*. The y-axis is the importance of the itemsets. The prediction models from left to right are: the independence model, the Gaussian model, prediction based on all sub-itemsets, the best tree model, and the best model found by the greedy algorithm.

In [3] we propose a new class of co-occurrence patterns: trees. The idea is to search for hierarchies of general and more specific attributes. The novelty in our approach is that we start from unordered data, and by using frequent pattern mining techniques infer hierarchical orders from data using a specific pattern scoring function.

Tree pattern mining has interesting applications in domains, such as text mining, where such tree patterns may reflect interesting co-occurrences between usages of terms. Figure 17.2 shows examples of such patterns discovered form a real text data sets of scientific research abstracts.

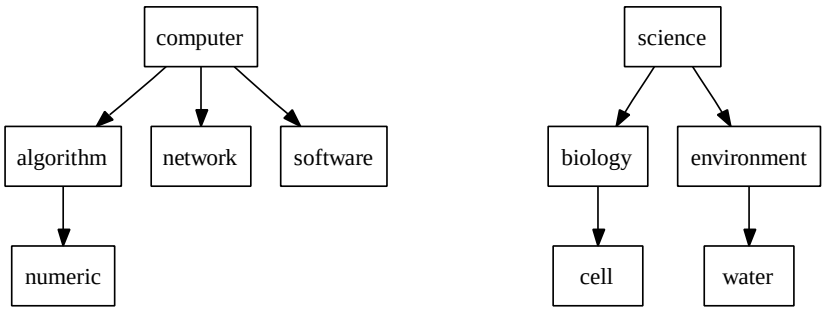


Figure 17.2: Two frequent trees patterns discovered from a text corpus of scientific research abstracts.

In [4] we study the use of entropy as a scoring function for frequent patterns. Using entropy we define low-entropy sets, a more general and expressive pattern class that of frequent sets. We show that entropy has many desired properties, such as the basic monotonicity, that allows to use the levelwise approach to efficiently discover all low-entropy patterns given some entropy threshold. Furthermore, we use entropy to defining a new tree pattern class, low-entropy trees, which can be seen as a probabilistic variant of the frequent tree pattern class defined in [3].

References

[1] Nikolaj Tatti. Maximum entropy based significance of itemsets. In *Proceedings of Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 312–321, 2007.

[2] Nikolaj Tatti. Maximum entropy based significance of itemsets. *Knowledge and Information Systems*. In press.

[3] H. Heikinheimo, H. Mannila, and J. K. Seppänen. Finding trees from unordered 0-1 data. In *Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 175–186, 2006.

[4] H. Heikinheimo, E. Hinkkanen, H. Mannila, T. Mielikäinen, and J. K. Seppänen. Finding low-entropy sets and trees from binary data. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 350–359, 2007.

Using patterns

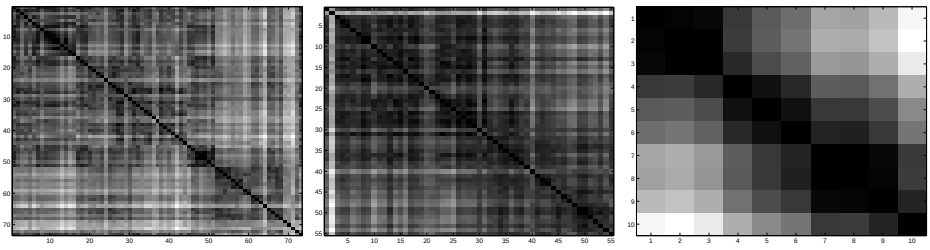
Nikolaj Tatti

A single itemset is a local pattern. It only represents a part of transactions and the information contained within the itemset is limited. However, a group of itemset can be very descriptive. In fact, if the group is large enough it can identify the whole dataset. While this is rarely the case in practice, we can think that a smaller group of itemsets captures the essential information of the original data. One of our research topic is how we can use itemsets in various tasks as a surrogate for data.

The theoretical part of these studies is in itself interesting and important, but there is also a practical point of view when we consider these scenarios in the context of privacy-preserving data mining. A group of itemsets can be viewed as a sanitized version of the actual data and the studies describe how we can work with itemsets without having the actual data.

In [1,2] we study how we can use these sets for predicting unknown itemsets. We show that this is an NP-hard problem but with certain assumptions we can ease the computational burden. The problem reduces to a linear program with an exponential number of variables. By applying the ideas from the theory of Markov Random Fields we are able to reduce the number of attributes. We also point out that our reduction is optimal, that is, if we reduce any additional attributes, then there is a dataset for which the additional reduction will alter the prediction.

In [3] we study how we can use itemsets for defining the distance between two binary datasets. Nowadays, there is a particular need for this type of work, since the amount of information keeps growing and getting more complex. Hence, we need algorithms and theorems in which a whole a database is considered as one data point. Once we have defined a distance between databases, we can expose these to traditional data mining tools, such as, clustering and visualization. We base the definition of our distance on a geometrical intuition. We show that we can derive the same distance from a different sets of axioms, hence giving the distance a strong theoretical support. We also described an efficient method for computing the distance: The distance is an Euclidean distance between the frequencies of certain parity formulae. We apply the distance for several datasets and demonstrate that the distance produces meaningful and interpretable results (see Figure 17.3).



(a) Books of Bible (b) Inaugural addresses of U.S. Presidents (c) Abstracts of NSF awards

Figure 17.3: Distance matrices for various datasets. We see a temporal behavior in Figures 17.3(b)–17.3(c) and a division between the Old Testament and the New Testament in Figure 17.3(a).

References

[1] Nikolaj Tatti. Computational complexity of queries based on itemsets. *Information Processing Letters*, pages 183–187, June 2006.

[2] Nikolaj Tatti. Safe projections of binary data sets. *Acta Informatica*, 42(8–9):617–638, April 2006.

[3] Nikolaj Tatti. Distances between data sets based on summary statistics. *Journal of Machine Learning Research*, 8:131–154, Jan 2007.

From models to patterns

Jaakko Hollmén, Jarkko Tikka, Samuel Myllykangas

In the context of bioinformatics, we are interested in describing DNA amplification patterns recorded as 0–1 data with compact and understandable descriptions. To understand the coarse structure of the amplifications (mutation patterns in the physical chromosome), we applied probabilistic clustering of 0-1 data with a finite mixture of multivariate Bernoulli distributions [3]. Model selection was performed with cross-validation based methods.

Clustering the data with the finite mixture model achieves a good clustering with six component distributions, depicted in the Figure 17.4. In [1], we have developed a method to describe the clusters with compact and understandable descriptions by extracting maximal frequent itemsets and transforming them to the nomenclature used to describe chromosomal areas.

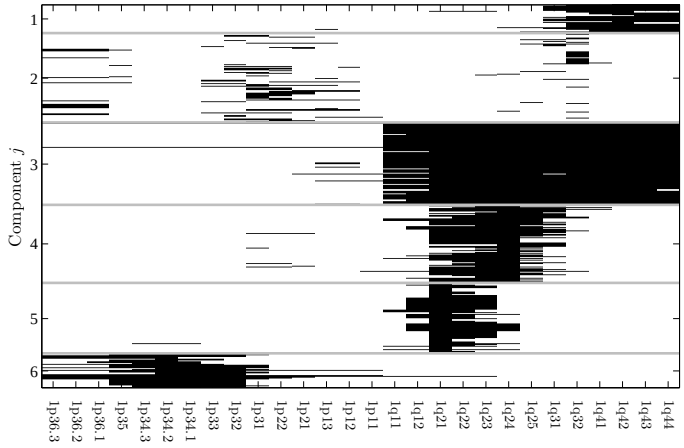


Figure 17.4: Clustered data from the human chromosome 1. Chromosomal locations are on the x-axis (marked under the figure) and the rows represent the clustered amplification patterns. Black areas are the DNA copy number amplifications. Compact and understandable descriptions are extracted from the cluster-specific data.

In a subsequent paper, the preceding results are reported in the problem of cancer classification [2]. We have clustered about 4500 cancer patients, one chromosome at the time and identified in total 111 amplification patterns in general. We investigated the associations of the amplification patterns with background factors of cancer types in order to underline the importance of a specific mutation in a particular cancer type.

References

[1] Jaakko Hollmén and Jarkko Tikka. Compact and understandable descriptions of mixture of Bernoulli distributions. In M.R. Berthold, J. Shawe-Taylor, and N. Lavrač, editors, *Proceedings of the 7th International Symposium on Intelligent Data Analysis (IDA 2007)*, volume 4723 of *Lecture Notes in Computer Science*, pages 1–12, Ljubljana, Slovenia, 2007. Springer-Verlag.

[2] Samuel Myllykangas, Jarkko Tikka, Tom Böhling, Sakari Knuutila, and Jaakko Hollmén. Classification of human cancers based on DNA copy number amplification patterns. Manuscript.

[3] Jarkko Tikka, Jaakko Hollmén, and Samuel Myllykangas. Mixture modeling of DNA copy number amplification patterns in cancer. In Francisco Sandoval, Alberto Prieto, Joan Cabestany, and Manuel Graña, editors, *Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN 2007)*, volume 4507 of *Lecture Notes in Computer Science*, pages 972–979, San Sebastián, Spain, 2007. Springer-Verlag.

17.2 Data mining theory

Cross-mining

Gemma C. Garriga, Hannes Heikinheimo, Jouni K. Seppänen

Most frequent pattern mining studies consider data sets with exclusively binary attributes. However, many real-world data sets have not only binary attributes but also numerical ones. As an example, consider movie recommendation systems: the binary data corresponds to movies rated by users, and the numerical data to their demographic statistics. A key problem is to segment users into similar groups according to their movie liking, but it is also in the interest of the service provider to produce descriptions of the groups for marketing purposes. Other such domains are for instance ecological data mining applications, where numerical environmental variates, such as rainfall or temperature, affect the binary occurrence (coexistence) of species across spatial locations.

In [1] we suggest a method for relating itemsets consisting of binary attributes to numerical attributes in the data. Our approach can be seen either as using the numerical attributes to measure the interestingness of the itemsets consisting of binary attributes, or, from another viewpoint, as using the itemsets to assess the interestingness of clusters or other local models found by mining the numerical attributes. Computing such models turns out to be computationally challenging, however, approximable within a constant factor using a simple greedy algorithm. Experiments show using biogeographical data that the algorithm can capture interesting patterns that would not have been found using either itemset mining or clustering alone.

References

[1] G. C. Garriga, H. Heikinheimo, and J. K. Seppänen. Cross-mining binary and numerical attributes. In *International Conference on Data Mining*, pages 481–486, 2007.

An approximation ratio for biclustering

Kai Puolamäki, Sami Hanhijärvi, Gemma C. Garriga

The problem of biclustering consists of the simultaneous clustering of rows and columns of a matrix such that each of the submatrices induced by a pair of row and column clusters is as uniform as possible. We have approximate the optimal biclustering by applying one-way clustering algorithms independently on the rows and on the columns of the input matrix. We have shown that such a solution yields a worst-case approximation ratio of $1 + \sqrt{2}$ under L_1-norm for 0–1 valued matrices, and of 2 under L_2-norm for real valued matrices.

Given a data matrix X, an optimal biclustering is a partition of rows and columns $\mathcal{R}$ and $\mathcal{C}$ into K_r and K_c partitions such that the cost

$$L = \sum_{R \in \mathcal{R}} \sum_{C \in \mathcal{C}} \mathcal{V}(X(R, C)),$$

is minimized, where we have used $\mathcal{V}(X(R, C))$ to denote the dissimilarity of the submatrix of X defined by the set of rows R and columns C.

Finding highly homogeneous biclusters has important applications for example in biological data analysis, where a bicluster may, for example, correspond to an activation pattern common to a group of genes only under specific experimental conditions.

We show that a straightforward algorithm, where a normal one way clustering algorithm is applied both to the rows and columns of the matrix. The scheme for approximating the optimal biclustering is defined as follows.

Input: matrix X, number of row clusters K_r, number of column clusters K_c

$\mathcal{R} = \text{kcluster}(X, K_r)$
$\mathcal{C} = \text{kcluster}(X^T, K_c)$

Output: a set of biclusters $X(R, C)$, for each $R \in \mathcal{R}$, $C \in \mathcal{C}$

---

kcluster(X, K) is a normal one way clustering algorithm, with a proven approximation ratio, which partitions the rows of matrix X into K clusters. This simple scheme gives an approximation ratio of $1 + \sqrt{2}$ for L_1 norm and 2 for L_2 norm, multiplied by the approximation ratio of the one-way clustering algorithm kcluster.

Our contribution shows that in many practical applications of biclustering, it may be sufficient to use a more straightforward standard clustering of rows and columns instead of applying heuristic algorithms without performance guarantees.

References

[1] Kai Puolamäki, Sami Hanhijärvi, Gemma C. Garriga. An Approximation Ratio for Biclustering. Publications in Computer and Information Science E13, arXiv:0712.2682v1. Accepted for publication in Information Processing Letters.

The cost of learning directed cuts

Gemma C. Garriga

Classifying vertices in digraphs is an important machine learning setting with many applications. We consider learning problems on digraphs with three characteristic properties: (i) The target concept corresponds to a directed cut; (ii) the total cost of finding the cut has to be bounded a priori; and (iii) the target concept may change due to a hidden context.

Recent learning theoretical work has concentrated on performance guarantees that depend on the complexity of the target function or at least on strong assumptions about the target function. In [1] we propose performance guarantees to learn a cut in a directed graph that only make natural assumptions about the target concept and that otherwise depend only on properties of the unlabelled training and test data. This is in contrast to related work on learning small cuts in undirected graphs where usually the size of the concept cut is taken as a (concept-dependent) learning parameter The first observation in [1] is that for learning directed cuts we can achieve tight, concept-independent guarantees based on a fixed size of the minimum path cover. We establish logarithmic performance guarantees for online learning, active learning, and PAC learning. We furthermore show which algorithms and results carry over to learning intersections of monotone with antimonotone concepts. An important contribution concerns learning algorithms able to cope with concept drift due to hidden changes in the context, i.e., the concept depends on an unobservable variable that can change over time. Worst case guarantees in this setting are related to adverserial learning.

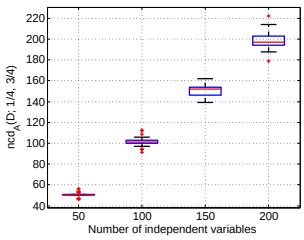
References

[1] T. Gärtner and G. C.Garriga. The Cost of Learning Directed Cuts. In *European Conference on Machine Learning (ECML)*, pages 152–163, 2007.

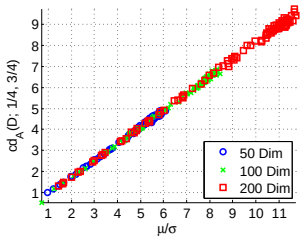
Dimensionality of data

Nikolaj Tatti and Heikki Mannila

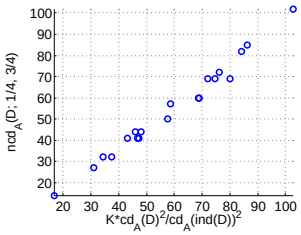
In [1] we study the intrinsic dimensionality of binary data. The idea of intrinsic dimension is rather important, since even though we may have very high-dimensional dataset, it may also possess a great amount of structure, hence its actual dimension is much smaller than the the number of attributes it is represented with. The concept of intrinsic dimension for binary data is a relatively new idea and it is an active topic in the data mining community. We apply fractal dimension — A concept that has strong theory base and has many applications with real-number datasets. We show that the fractal dimension has many interesting properties, however, the fractal dimension has some problems that are typical only to binary data. The value of the fractal dimension tend to be small for sparse binary data. Hence we tailor a new concept called the normalized fractal dimension. This dimension does not depend on the sparsity of data. In our experiments we study various properties of the dimension and compare it against several base measures.



(a) Normalized dimension, generated data



(b) Fractal dimension, generated data



(c) Normalized dimension, *20 Newsgroups*

Figure 17.5: Dimensions for various datasets. In Figure 17.5(a) we see that the normalized dimension is essentially the number of attributes if the attributes are independent. In Figures 17.5(b)–17.5(c) dimensions are plotted as functions of their estimates.

References

[1] Nikolaj Tatti, Taneli Mielikäinen, Aristides Gionis, and Heikki Mannila. What is the dimension of your binary data. In *Proceedings of Sixth IEEE International Conference on Data Mining (ICDM 2006)*, pages 603–612, 2006.

Miscellaneous problems

In [3] we introduced the discrete basis problem, a discrete version of matrix decomposition problems, and analyzed its complexity.

One of the problems in clustering is that different parameter settings can give different results. A possible solution is to use clustering aggregation, i.e., given a set of clusterings, combine them to a single clustering that agrees as well as possible with the given clusterings [2] While this problem is NP-hard, simple algorithms have a constant approximation ratio, and perform very well in practice.

Sampling in different forms is a strong technique in data analysis. The paper [1] formulates the problem of sampling hidden databases, i.e., getting unbiased samples from data sources that can be accessed only via queries. We show that the problem can, in several cases, be solved, and give simple yet powerful algorithms for the task.

References

[1] A. Dasgupta, G. Das, and H. Mannila. A Random Walk Approach to Sampling Hidden Databases. Proceedings of the 2007 ACM SIGMOD international conference on Management of Data (SIGMOD 2007), p. 629–640.

[2] A. Gionis, H. Mannila, P. Tsaparas. Clustering Aggregation (long version). ACM Transactions on Knowledge Discovery from Data, 1, 1 (2007),

[3] P. Miettinen, T. Mielikäinen, A. Gionis, G. Das, H. Mannila. The Discrete Basis Problem. 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD) 2006, p. 335–346. PKDD Best Paper.

17.3 Analyzing ordered data

Heikki Mannila, Kai Puolamäki, Antti Ukkonen

We have continued our work on developing data analysis algorithms both for finding and analyzing orders. Two results are discussed in this section, while a third, related topic is considered in the section related to randomization techniques.

Finding bucket orders

Given a set of *pairwise preferences* over a finite set of items M, we have considered the problem of ordering the items so that the resulting order agrees as much as possible with the given preferences. This problem is motivated for example by biostratigraphy, where the task is to determine the age of sediments using the fossils they contain. We use the proposed method for finding a temporal order for a number of sites (geographical locations) where fossils have been found. In this case the pairwise preferences tell us for each pair of sites which one of them is more likely to be older, and hence to precede the other one in a temporal ordering of the sites. Methods for estimating these probabilities from data based on the fossil record are considered in [1,2].

Usual approaches to combining pairwise preferences to a global order over the items try to produce a total order. This means that given the result on the fossil discovery sites, we can say for every two sites which one of them is older. There are some problems related to this, however. First, a total order may be an incorrect model class to begin with. In the fossil application it is reasonable to assume that the sites belong to certain paleontological eras. If two sites belong to the same era, it may be very difficult even for a skilled expert in stratigraphy to determine which one of the sites is in fact older. Hence, a model where the sites are not totally ordered, but placed in classes that correspond to different temporal periods is a justified approach.

Second, the preferences are likely to contain noise in some form. It is possible that two items appear to be ordered in a certain way due to random chance. Distinguishing such items from those for which the ordering is more certain is not possible from a total order. However, if we use a model that can leave some pairs of sites unordered, we can avoid errors that have been introduced to the result due to noise in the input.

Our result is an algorithm that finds a *bucket order* given the pairwise preferences. The preferences are expressed as probabilities $P(u \prec v)$ for every $u, v \in M$. This is the probability of the item u to precede item v in a global order on the items. A bucket order is a total order with ties, i.e., a disjoint partition of M to k buckets together with a total order on the buckets. In the paleontology example this means that the very oldest sites are put in the first bucket, the second oldest in the second, and so on, with the youngest sites in the last bucket.

More precisely, if the pairwise probabilities are stored in a $|M| \times |M|$ matrix C, the algorithm tries to find another matrix B that has a direct interpretation as a bucket order, and that is a good approximation of the original matrix C. In practice the algorithm minimizes a cost function of the form $|C - B|_1$, i.e., the L_1 norm between C and B. This problem turns out to be NP-hard, but the randomized algorithm we propose has an expected constant factor approximation guarantee.

C_1	C_2
8 9 10 5 12 6	5 10 4 12 1 11
9 10 5 3 11 6	5 2 3 4 12 11
8 9 3 5 11 6	6 3 8 4 11 7
9 8 5 11 12 2	5 2 4 8 1 9
7 9 10 5 3 6	2 12 8 7 9 1
7 1 5 2 6 4	2 5 4 3 7 1
1 9 3 5 2 12	5 2 6 4 12 1
9 5 2 11 12 4	2 6 10 4 9 1
7 8 3 12 2 4	5 10 3 9 7 11
9 8 3 5 2 4	5 6 10 12 9 7
9 10 3 12 2 11	5 3 8 9 11 1
7 10 2 12 11 6	5 10 3 4 1 11
7 6 12 2 11 4	2 6 3 8 9 1
1 8 11 12 6 4	2 6 8 12 7 11
7 9 8 5 6 4	5 10 12 7 1 11

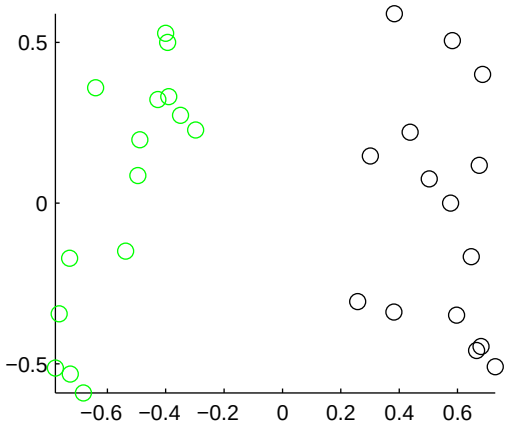


Figure 17.6: Visualizing sets of chains: On the left is a set of chains that can be divided to two groups, C_1 and C_2, based on the contents of the chains. On the right is a scatterplot where chains belonging to C_1 are indicated in green and chains belonging to C_2 appear in black. (Image from [2].)

Visualizing sets of chains

A *chain* is a totally ordered subset of a finite set of items M. For example, if M contains the title of every movie that came out in 2007, and we ask a person to rank those movies of M she has seen according to her preferences, the resulting order is a chain on M. If we ask a number of different people to rank the movies they have seen, we obtain a set of chains on the movie titles. This type of data can be used for instance to divide the respondents to a number of groups so that respondents in the same group have similar preferences. Market analysis and collaborative filtering are examples of practical applications of this approach.

We have considered different techniques for representing a set of chains as a two dimensional scatterplot where each chain appears as a single point. Our aim is to construct the visualization so that points associated to similar chains appear close to each other in the figure. Such visualizations can be useful for example for identifying structure in data or for manual classification of unseen data points.

The basic approach we take is to first map each chain of the input to a single point in a high dimensional euclidean space. Subsequently some dimension reduction method is applied to this set of points to obtain the final scatterplot. Our main contribution in this work is the development of two techniques for mapping chains to vector spaces. A central problem related to this is that comparing two chains that have no items in common is in general hard. It is possible that two chains should be mapped to points that are close to each other in the high dimensional space despite them having no common items. This happens for instance when the chains are generated by two different components; chains emitted by the same component should be placed closer to each other and away from those emitted by the other component.

Our first mapping is based on comparing each chain in the input with the other chains, and the second one maps the chains directly to points on the surface of a high dimensional hypersphere. While the first approach will place chains that have been generated by the same component close to each other, it can in practice be slow. The latter method is very

fast in comparison, but it does not attempt to recognize if two chains are emitted by the same component. In practice the two results seem to give similar results, however.

References

[1] Kai Puolamäki, Mikael Fortelius and Heikki Mannila. Seriation in Paleontological Data Using Markov Chain Monte Carlo Methods. In PLoS Computational Biology 2(2): e6, 2006.

[2] Antti Ukkonen. Visualizing Sets of Partial Rankings. In Michael R. Berthold, John Shawe-Taylor and Nada Lavrac, editors, *Advances in Intelligent Data Analysis: IDA 2007*, volume 4723 of *Lecture Notes in Computer Science*, pages 240–251. Springer, 2007.

[3] Antti Ukkonen. Algorithms for Finding Orders and Analyzing Sets of Chains. Ph.D Thesis, Department of Information and Computer Science, Helsinki University of Technology, 2008.

17.4 Randomization methods in data analysis

Heikki Mannila and Antti Ukkonen

Swap randomization of 0–1 data

Determining whether data analysis results could be the result of chance is obviously an important task. Traditional statistical significance testing methods are not very suitable for assessing the results of complex data mining operations, as the distributional assumptions behind the traditional methods are typically always false. We have been considering randomization-based methods in different contexts.

In [2] we consider a simple randomization technique for producing random datasets that have the same row and column margins with the given dataset. Then one can test the significance of a data mining result by computing the results of interest on the randomized instances and comparing them against the results on the actual data. This randomization technique can be used to assess the results of many different types of data mining algorithms, such as frequent sets, clustering, and rankings. To generate random datasets with given mar- gins, we use variations of a Markov chain approach, which is based on a simple swap operation. We give theoretical results on the efficiency of different randomization methods, and apply the swap randomization method to several well- known datasets. Our results indicate that for some datasets the structure discovered by the data mining algorithms is a random artifact, while for other datasets the discovered structure conveys meaningful information.

A similar approach is used in [1] for comparing segmentations.

Randomization of chains

A *chain* is a total order on some subset of a finite set of items M. See Section 17.3 for a more thorough description and example of a set of chains. Here we consider an algorithm for creating random sets of chains that share a number of statistics with a given set of chains.

In case of 0–1 data preserving the row and column sums is of interest as argued above. With chains we want to maintain a number of different statistics. When running a data analysis algorithm on a set of chains, we are essentially investigating the rankings, i.e., the results we obtain should be a consequence of the *ordering information* present in D. This is only one property of the input. Others are *the number of chains in the set*, *distribution of the lengths of the chains*, *frequencies of all itemsets* (when each chain is viewed as a set of items), and *the number of times the item u precedes the item v for all $u, v \in M$*.

The first property simply states that the random data sets should be of the same size as the original one. This is a very intuitive requirement. Maintaining the second property rules out the possibility that the found results are somehow caused only by the length distribution of the chains in the input. Likewise, maintaining the itemsets should rule out the possibility that the result is not a consequence of the rankings, but simply the co-occurrences of the items. Finally, maintaining the pairwise frequencies is analoguous to maintaining the mean of a set of real valued vectors.

The method we propose generates a random data set that is equivalent to the given data set if only the above properties are considered. It is a Markov Chain Monte Carlo algorithm that starts from the given set of chains and makes small local modifications to the chains that are guaranteed to preserve all of the properties above. These modifications affect two chains at a time, transposing two adjacent items in both. In general the algorithm will be run until it reaches a state that is no longer correlated with the initial state. This final state is selected as the random data set. This process is repeated until enough random data sets have been sampled.

References

[1] Niina Haiminen, Heikki Mannila, Evimaria Terzi. Comparing segmentations by applying randomization techniques. BMC Bioinformatics 2007, 8:171 (23 May 2007).

[2] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing Data Mining Results via Swap Randomization. ACM Transactions on Knowledge Discovery from Data (TKDD), Volume 1 , Issue 3 (December 2007) Article No. 14.

17.5 Applications

Ecological applications

Hannes Heikinheimo and Heikki Mannila

Many central ecological questions are related to understanding how different species communities form and what are the kind of effects climate has them. Related to this, we have applied data mining methods to study the spatial distributions of European land mammal fauna and their relationship to the environment [1]. Using clustering techniques we found that the mammalian species divide naturally into clusters, which are highly connected spatially, and that the clusters reflect major physiographic and environmental features and differ significantly in the values of basic climate variables (see Figure 17.7).

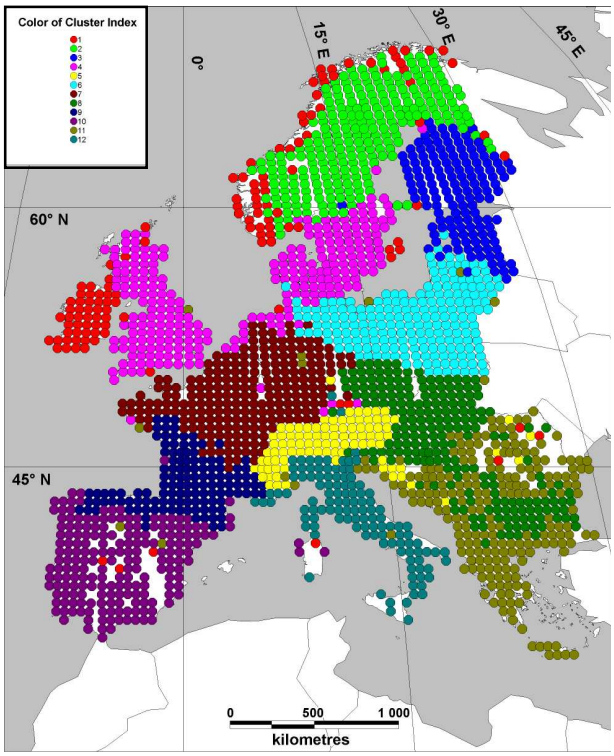


Figure 17.7: Spatial clustering of Europe using species occurrence data of 124 European land mammal species.

Our palette of applications includes a wide variety of topics. For example, in [5] we consider the problems in analyzing datasets arising in the study of linguistic change: the key issue is the small sample size for any particular period of time. The article [4] looks at haplotyping (a genetic data analysis problem) by using string models of variable length.

The ecological theme is strongly present in [2], where we study the computational properties of nestedness, a concept arising from ecology, and its generalization, segmented nestedness. These concepts turn out to be quite useful also for other applications, and their combinatorial and algorithmic properties are challenging.

References

[1] H. Heikinheimo, M. Fortelius, J. Eronen, and H. Mannila. Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography*, 34(6):1053–1064, 2007.

[2] H. Mannila, E. Terzi. Nestedness and segmented nestedness. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2007), p. 480–489.

[3] N. Haiminen, H. Mannila. Discovering isochores by least-squares optimal segmentation. Gene 394 (Issues 1–2), 2007, pp. 53–60 (1 June 2007).

[4] N. Landwehr, T. Mielikäinen, L. Eronen, H. Toivonen, H. Mannila. Constrained hidden Markov models for population-based haplotyping. BMC Bioinformatics 2007, 8(Suppl 2):S9.

[5] A. Hinneburg, H. Mannila, S. Kaislaniemi, T. Nevalainen and H. Raumolin-Brunberg. How to Handle Small Samples: Bootstrap and Bayesian Methods in the Analysis of Linguistic Change. Literary and Linguistic Computing 22, 2 (June 2007) 137–150; doi: 10.1093/llc/fqm006

Seriation of paleontological data

Kai Puolamäki and Heikki Mannila

Seriation, the task of temporal ordering of fossil occurrences by numerical methods, and correlation, the task of determining temporal equivalence, are fundamental problems in paleontology. With the increasing use of large databases of fossil occurrences in paleontological research, the need is increasing for seriation methods that can be used on data with limited or disparate age information.

We have developed a simple probabilistic model of site ordering and taxon occurrences. As there can be several parameter settings that have about equally good fit with the data, we have developed Bayesian Markov chain Monte Carlo methods to obtain a sample of parameter values describing the data. As an example, the method is applied to a dataset on Cenozoic mammals.

The orderings produced by the method agree well with the orderings of the sites with known geochronologic ages.

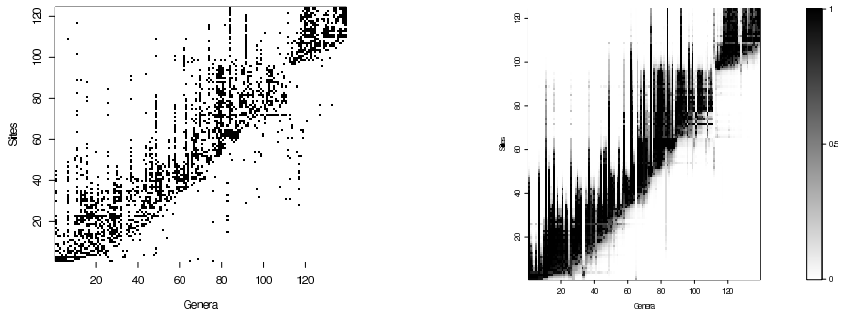


Figure 17.8: The original paleontological data matrix, where the rows correspond to the find sites and columns to the genera, and black that a genus has been found from the find site (left), as well as the probability that the genus existed during the time period of a given site (right).

References

[1] Kai Puolamäki, Mikael Fortelius, Heikki Mannila. Seriation in Paleontological Data Using Markov Chain Monte Carlo Methods. PLoS Computational Biology 2(2): e6, 2006.

17.6 Segmentation

Heikki Mannila and Robert Gwadera

For sequential data, segmentation is the counterpart of clustering. We have in recent years studied different aspects of segmentation, both theory and practice. The basic segmentation problem can be solved in polynomial time by using dynamic programming, and there are several interesting variants for study.

Segmental prediction is applicable in situations where the phenomenon of interest is governed by different models at different points of time. Such phenomena occur naturally in, e.g., atmospheric data, where for example winter and summer conditions for aerosol formation differ qualitatively and quantitatively. In [1] we studied the combination of dynamic programming and facility location approaches to obtain a small set of recurring models to be used in prediction.

A biological application of dynamic programming for the discovery of isochore structure is given in [3], while [4] looks at the randomization models needed for comparing segmentations. In [2] we study the segmentation of models of different depth for segmenting strings, especially DNA.

References

[1] S. Hyvönen, A. Gionis, H. Mannila. Recurrent predictive models for sequence segmentation. Advances in Intelligent Data Analysis VII (IDA 2007), p. 195–206.

[2] R. Gwadera, A. Gionis, H. Mannila. Optimal Segmentation using Tree Models. 2006 IEEE International Conference on Data Mining, p. 244–253, 2006

[3] N. Haiminen, H. Mannila. Discovering isochores by least-squares optimal segmentation. Gene 394 (Issues 1–2), 2007, pp. 53–60 (1 June 2007).

[4] Niina Haiminen, Heikki Mannila, Evimaria Terzi. Comparing segmentations by applying randomization techniques. BMC Bioinformatics 2007, 8:171 (23 May 2007).

Publications of the From Data to Knowledge Research Unit

[1] E. Bingham, A. Gionis, N. Haiminen, H. Hiisilä, H. Mannila, and E. Terzi. Segmentation and dimensionality reduction. In *2006 SIAM Conference on Data Mining 2006*, pages 372–383, 2006.

[2] A. Dasgupta, G. Das, and H. Mannila. A random walk approach to sampling hidden databases. In *Proc. of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD 2007)*, pages 629–640, 2007.

[3] G. C. Garriga, H. Heikinheimo, and J. K. Seppänen. Cross-mining binary and numerical attributes. In *Proc. of IEEE International Conference on Data Mining (ICDM)*, pages 481–486, 2007.

[4] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(3):14, 2007.

[5] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. In *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2006*, pages 167–176, 2006.

[6] A. Gionis, H. Mannila, K. Puolamäki, and A. Ukkonen. Algorithms for discovering bucket orders from data. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 561–566, 2006.

[7] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation (long version). *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[8] R. Gupta, S. Ruosaari, S. Kulathinal, J. Hollmén, and P. Auvinen. Microarray image segmentation using additional dye - an experimental study. *Molecular and Cellular Probes*, (5-6):321–328, 2007.

[9] R. Gwadera, A. Gionis, and H. Mannila. Optimal segmentation using tree models. In *2006 IEEE International Conference on Data Mining 2006*, pages 244–253, 2006.

[10] R. Gwadera, J. Toivola, and J. Hollmén. Segmenting multi-attribute sequences using dynamic Bayesian networks. In *Proceedings of The Seventh IEEE International Conference on Data Mining - Workshops (ICDM Workshops 2007)*, pages 465–470. IEEE Computer Society, 2007.

[11] N. Haiminen and H. Mannila. Discovering isochores by least-squares optimal segmentation. *Gene*, 394(1-2):53–60, 2007.

[12] N. Haiminen, H. Mannila, and E. Terzi. Comparing segmentations by applying randomization techniques. *BMC Bioinformatics*, 8(171 (23 May 2007)), 2007.

[13] D. R. Hardoon, J. Shawe-Taylor, A. Ajanki, K. Puolamäki, and S. Kaski. Information retrieval by inferring implicit queries from eye movements. In M. Meila and X. Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

[14] H. Heikinheimo, M. Fortelius, J. Eronen, and H. Mannila. Biogeography of european land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography*, 34(6):1053–1064, 2007.

[15] H. Heikinheimo, E. Hinkkanen, H. Mannila, T. Mielikäinen, and J. K. Seppänen. Finding low-entropy sets and trees from binary data. In *Proc. of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 350–359, 2007.

[16] H. Heikinheimo, H. Mannila, and J. K. Seppänen. Finding trees from unordered 0-1 data. In J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, editors, *Knowledge Discovery in Databases: PKDD 2006*, pages 175–186. Springer, 2006.

[17] A. Hinneburg, H. Mannila, S. Kaislaniemi, T. Nevalainen, and H. Raumolin-Brunberg. How to handle small samples: Bootstrap and Bayesian methods in the analysis of linguistic change. *Literary and Linguistic Computing*, 22(2):137–150, 2007.

[18] J. Hollmén. Model selection and estimation via subjective user preferences. In V. Corruble, M. Takeda, and E. Suzuki, editors, *Proceedings of The Tenth International Conference on Discovery Science (DS 2007)*, pages 259–263, Sendai, 2007. Springer-Verlag.

[19] J. Hollmén and J. Tikka. Compact and understandable descriptions of mixture of Bernoulli distributions. In M. Berthold, J. Shawe-Taylor, and N. Lavrac, editors, *Proceedings of the 7th International Symposium on Intelligent Data Analysis (IDA 2007)*, pages 1–12, Ljubljana, 2007. Springer-Verlag.

[20] S. Hyvönen, A. Gionis, and H. Mannila. Recurrent predictive models for sequence segmentation. In *Advances in Intelligent Data Analysis VII (IDA 2007) 2007*, pages 195–206, 2007.

[21] H. Keski-Säntti, T. Atula, J. Tikka, J. Hollmén, A. A. Mäkitie, and I. Leivo. Predictive value of histopathologic parameters in early squamous cell carcinoma of oral tongue. *Oral Oncology*, 43(10):1007–1013, 2007.

[22] M. Korpela and J. Hollmén. Extending an algorithm for clustering gene expression time series. In J. Rousu, S. Kaski, and E. Ukkonen, editors, *Probabilistic Modeling and Machine Learning in Structural and Systems Biology 2006*, pages 120–124, Helsinki, 2006.

[23] N. Landwehr, T. Mielikäinen, L. Eronen, H. Toivonen, and H. Mannila. Constrained hidden markov models for population-based haplotyping. *BMC Bioinformatics*, 8 (Suppl 2):1–9, 2007.

[24] N. Landwehr, T. Mielikäinen, L. Eronen, H. Toivonen, and H. Mannila. Constrained hidden markov models for population-based haplotyping. In *Probabilistic Modeling and Machine Learning in Structural and Systems Biology (PMSB 2006) 2006*, 2006.

[25] P. Lindholm, P. Nymark, H. Wikman, K. Salmenkivi, A. Nicholson, M. V. Korpela, S. Kaski, S. Ruosaari, J. Hollmen, E. Vanhala, A. Karjalainen, S. Anttila, V. Kinnula, and S. Knuutila. Asbestos-associated malignancies in the lung and pleura show distinct genetic aberrations. *Lung cancer*, 54:15, 2006.

[26] S. Luyssaert, I. Janssens, M. Sulkava, D. Papale, A. Dolman, M. Reichstein, T. Suni, J. Hollmén, T. Vesala, D. Lousteau, B. Law, and E. Moors. Photosynthesis drives interannual variability in net carbon-exchange of pine forests at different latitudes. In *Proceedings of the Open Science Conference on the GHG Cycle in the Northern Hemisphere*, pages 86–87, Jena, Germany, 2006. CarboEurope, NitroEurope, CarboOcean, and Global Carbon Project.

[27] S. Luyssaert, I. A. Janssens, M. Sulkava, D. Papale, A. J. Dolman, M. Reichstein, J. Hollmén, J. G. Martin, T. Suni, T. Vesala, D. Lousteau, B. E. Law, and E. J. Moors. Photosynthesis drives anomalies in net carbon-exchange of pine forests at different latitudes. *Global Change Biology*, 13(10):2110–2127, 2007.

[28] S. Luyssaert, M. Sulkava, H. Raitio, J. Hollmén, and P. Merilä. Is n and s deposition altering the mineral nutrient composition of norway spruce and scots pine needles in finland? In J. Eichhorn, editor, *Proceedings of Symposium: Forests in a Changing Environment - Results of 20 years ICP Forests Monitoring*, pages 80–81, Göttingen, Germany, 2006. ICP Forests, European Commission, Nordwestdeutsche Versuchsanstalt.

[29] H. Mannila. The role of information technology for systems biology. in systems biology: A grand challenge for europe, esf 2007, p. 21-23., 2007.

[30] H. Mannila and E. Terzi. Nestedness and segmented nestedness. In *Proc. of 13th ACM SIGKDD International Conference on Knowledge*, pages 480–489, 2007.

[31] P. Miettinen, T. Mielikäinen, A. Gionis, G. Das, and H. Mannila. The discrete basis problem. In *10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD) 2006 2006*, 2006.

[32] S. Myllykangas, J. Himberg, T. Böhling, B. Nagy, J. Hollmén, and S. Knuutila. Dna copy number amplification profiling of human neoplasms. *Oncogene*, 25(55):7324–7332, 2006.

[33] P. Nymark, P. M. Lindholm, M. V. Korpela, L. Lahti, S. Ruosaari, S. Kaski, J. Hollmen, S. Anttila, V. L. Kinnula, and S. Knuutila. Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines. *BMC Genomics*, 8:1–14, 2007.

[34] P. Nymark, H. Wikman, S. Ruosaari, J. Hollmén, E. Vanhala, A. Karjalainen, S. Anttila, and S. Knuutila. Identification of specific gene copy number changes in asbestos-related lung cancer. *Cancer Research*, 66(11):5737–5743, 2006.

[35] K. Puolamäki, M. Fortelius, and H. Mannila. Seriation in paleontological data using markov chain monte carlo methods. *PLoS Computational Biology*, 2(2):26, 2006.

[36] K. Puolamäki, M. Fortelius, and H. Mannila. Seriation in paleontological data using markov chain monte carlo methods, 2006.

[37] K. Puolamäki, S. Hanhijärvi, and G. C. Garriga. An approximation ratio for biclustering. Technical Report Report E13, Espoo, Finland, 2007.

[38] K. Puolamäki and S. Kaski. *Proceedings of the NIPS 2005 Workshop on Machine Learning for Implicit Feedback and User Modeling*. Espoo, 2006.

[39] K. Puolamäki, J. Salojärvi, E. Savia, and S. Kaski. Discriminative mcmc. Technical Report Report E1, Helsinki University of Technology, Espoo, Finland, 2006.

[40] A. Rasinen, J. Hollmén, and H. Mannila. Analysis of linux evolution using aligned source code segments. In N. Lavrac, L. Todorovski, and K. Jantke, editors, *Proceedings of the Ninth International Conference on Discovery Science*, pages 209–218. Springer-Verlag, 2006.

[41] E. Salmela, O. Taskinen, J. K. Seppänen, P. Sistonen, M. J. Daly, P. Lahermo, M.-L. Savontaus, and J. Kere. Subpopulation difference scanning: a strategy for exclusion mapping of susceptibility genes. *Journal of medical genetics*, 43:590–597, 2006.

[42] M. Sulkava, S. Luyssaert, P. Rautio, I. A. Janssens, and J. Hollmén. Modeling the effects of varying data quality on trend detection in environmental monitoring. *Ecological Informatics*, 2(2):167–176, 2007.

[43] M. Sulkava, H. Mäkinen, P. Nöjd, and J. Hollmén. CUSUM charts for detecting onset and cessation of xylem formation based on automated dendrometer data. In I. Horová and J. Hrebícek, editors, *TIES 2007 - 18th annual meeting of the International Environmetrics Society*, page 111, Mikulov, 2007. Masaryk University.

[44] M. Sulkava, J. Tikka, and J. Hollmén. Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. *Ecological Modelling*, 191:118–130, 2006.

[45] N. Tatti. Computational complexity of queries based on itemsets. *Information Processing Letters*, pages 183–187, 2006.

[46] N. Tatti. Safe projections of binary data sets. *Acta Informatica*, 42(8-9):617–638, 2006.

[47] N. Tatti. Distances between data sets based on summary statistics. *Journal of Machine Learning Research*, 8:131–154, 2007.

[48] N. Tatti. Maximum entropy based significance of itemsets. In *Proc. of Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 312–321, 2007.

[49] N. Tatti, T. Mielikäinen, A. Gionis, and H. Mannila. What is the dimension of your binary data? In *2006 IEEE International Conference on Data Mining 2006*, pages 603–612, 2006.

[50] J. Tikka and J. Hollmén. Long-term prediction of time series using a parsimonious set of inputs and LS-SVM. In A. Lendasse, editor, *Proceedings of the First European Symposium on Time Series Prediction (ESTSP 2007)*, pages 87–96, Espoo, 2007.

[51] J. Tikka and J. Hollmén. A sequential input selection algorithm for long-term prediction of time series. *Neurocomputing*, 2007.

[52] J. Tikka, J. Hollmén, and S. Myllykangas. Mixture modeling of DNA copy number amplification patterns in cancer. In F. Sandoval, A. Prieto, J. Cabestany, and M. Grana, editors, *Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN 2007)*, pages 972–979, San Sebastián, 2007.

[53] J. Tikka, A. Lendasse, and J. Hollmén. Analysis of fast input selection: Application in time series prediction. In *International Conference on Artificial Neural Networks (ICANN) 2006*, pages 161–170, 2006.

[54] A. Ukkonen. Visualizing sets of partial rankings. In *Proc. of Advances in Intelligent Data Analysis VII, 7th International Symposium on Intelligent Data Analysis (IDA 2007)*, pages 240–251, 2007.

[55] A. Ukkonen and H. Mannila. Finding outlying items in sets of partial rankings. In *Proc. of PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 265–276, 2007.

[56] H. Wikman, S. Ruosaari, P. Nymark, V. K. Sarhadi, J. Saharinen, E. Vanhala, A. Karjalainen, J. Hollmén, S. Knuutila, and S. Anttila. Gene expression and copy number profiling suggests the importance of allelic imbalance in 19p in asbestos-associated lung cancer. *Oncogene*, 26(32):4730–4737, 2007.